Introduction to Power Analysis

A Guide to G*Power, jamovi, and Superpower

James Bartlett (james.bartlett@glasgow.ac.uk)

@JamesEBartlett

Version 1.4 (01/02/2022)

1. Installing the software used in this guide

There are different tools you can use to calculate power. We are mainly going to use G*Power, but this will not work on Apple Macs running the Catalina OS or newer since it disables 32 bit software. Some of the tests will also be demonstrated in jamovi, but it currently covers a limited suite of statistical tests. Both of these software are limited in their ability to calculate power for factorial designs, so we will use the Shiny App created by Lakens and Caldwell (2021) for this. Hopefully, this will act as a trojan horse to get you thinking about running power analyses in R as it is the most flexible approach, but you ideally need a firm grasp of using R first.

G*Power is unfortunately no longer in development, with the last update being in July 2017. You can download G*Power on this page. Under the heading “download” click on the appropriate version for whether you have a Windows or Mac computer.

If you need jamovi, you can download it from this page. Install the solid version to be the most stable, and once opened, you will need to click Module (top right) > jamovi library > install jpower. This is an additional module you can add to your jamovi toolbar to calculate power for comparing two groups or two conditions.

2. Why should you perform a power analysis?

2.1. Principles of frequentist statistics

In order to utilise power analysis, it is important to understand the statistics we commonly use in psychology, known as frequentist or classical statistics. This is a theory of statistics where probability is assigned to long-run frequencies of observations, rather than assigning a likelihood to one particular event. This is the basis of where you get *p* values from. The formal definition of a *p* value is the probability of observing a result at least as extreme as the one observed, assuming the null hypothesis is true (Cohen, 1994). This means a small *p* value indicates the results are surprising if the null hypothesis is true and a large *p* value indicates the results are not very surprising if the null is true. The aim of this branch of statistics is to help you make decisions and limit the amount of errors you will make in the long-run (Neyman, 1977). There is a real emphasis on long-run, as the probabilities do not relate to individual cases or studies, but tell you the probability attached to the procedure if you repeated it many times.

There are two important concepts here: alpha and beta. Alpha is the probability of concluding there is an effect when there is not one (type I error or false positive). This is normally set at .05 (5%) and it is the threshold we look at for a significant effect. Setting alpha to .05 means we are willing to make a type I error 5% of the time in the long run. Beta is the probability of concluding there is not an effect when there really is one (type II error or false negative). This is normally

set at .20 (20%), which means we are willing to make a type II error 20% of the time in the long run. These values are commonly used in psychology, but you could change them. However, both values should ideally decrease rather than increase the number of errors you are willing to accept.

Power is the ability to detect an effect if there is one there to be found. In other words, “if an effect is a certain size, how likely are we to find it?” (Baguley, 2004: 73). Power relates to beta, as power is 1 - beta. Therefore, if we set beta to .2, we can expect to detect a particular effect size 80% of the time if we repeated the procedure over and over. Carefully designing an experiment in advance allows you to control the type I and type II error rate you would expect in the long run. However, these two concepts are not given much thought when designing experiments.

2.2. The problem of low statistical power

There is a long history of warnings about low power. One of the first articles was Cohen (1962) who found that the sample sizes used in articles only provided enough power to detect large effects (by Cohen’s guidelines, a standardised mean difference of 0.80 - See Section 2.3 below for an overview of effect sizes). The sample sizes were too small to reliably detect small to medium effects. This was also the case in the 1980s (Sedlmeier & Gigerenzer, 1989) and it is still a problem in contemporary research (Button et al., 2013). One reason for this is researchers often use “rules of thumb”, such as always including 20 participants per cell. However, this is not an effective strategy. Even experienced researchers overestimate the power provided by a given sample size and underestimate the number of participants required for a given effect size (Bakker et al., 2016).This shows you need to think carefully about power when you are designing an experiment.

The implications of low power is a waste of resources and a lack of progress. A study that is not sensitive to detect the effect of interest will just produce a non-significant finding more often than not. However, this does not mean there is no effect, just that your test was not sensitive enough to detect it. One analogy (paraphrased from this lecture by Richard Morey) to help understand this is trying to tell two pictures apart that are very blurry. You can try and squint, but you just cannot make out the details to compare them with any certainty. This is like trying to find a significant difference between groups in an underpowered study. There might be a difference, but you just do not have the sensitivity to differentiate the groups.

In order to design an experiment to be informative, it should be sufficiently powered to detect effects which you think are practically interesting (Morey & Lakens, 2016). This is called the *smallest effect size of interest*. Your test should be sensitive enough to avoid missing any values that you would find practically or theoretically interesting (Lakens, 2021). Fortunately, there is a way to calculate how many participants are required to provide a sufficient level of power known as power analysis.

In the simplest case, there is a direct relationship between statistical power, the effect size, alpha, and the sample size. This means that if you know three of these values, you can calculate the fourth. For more complicated types of analyses, you need some additional parameters, but we will tackle this as we come to it. The most common types of power analysis are *a priori* and sensitivity.

An *a priori* power analysis tells you how many participants are required to detect a given effect size. You are asking the question: “Given a known effect size, how many participants would I need to detect it with X% power?”. A sensitivity power analysis tells you what effect sizes your sample size is sensitive to detect. This time, you are asking the question: “Given a known sample size, what effect size could I detect with X% power?”.

Both of these types of power analysis can be important for designing a study and interpreting the results. If you need to calculate how many participants are required to detect a given effect, you can perform an *a priori* power analysis. If you know how many participants you have (for example you may have a limited population or did not conduct an *a priori* power analysis), you can perform a sensitivity power analysis to calculate which effect sizes your study is sensitive to detect.

Another type of power analysis you might come across is post-hoc. This provides you with the observed power given the sample size, effect size, and alpha. You can actually get software like SPSS to provide this in the output. However, this type of power analysis is not recommended as it fails to consider the long run aspect of these statistics. There is no probability attached to individual studies. You either observe an effect (significant *p* value) or you do not observe an effect (non-significant *p* value). I highly recommend ignoring this type of power analysis and focusing on *a priori* or sensitivity power analyses.

2.3. Effect sizes

In previous sections, I talked about effects and effect sizes, but you might not know what these are. When we use hypothesis testing, we are interested in things like distinguishing two groups or seeing if there is an association between two variables. The effect size is the actual difference between your two groups or the strength of the association between your two variables.

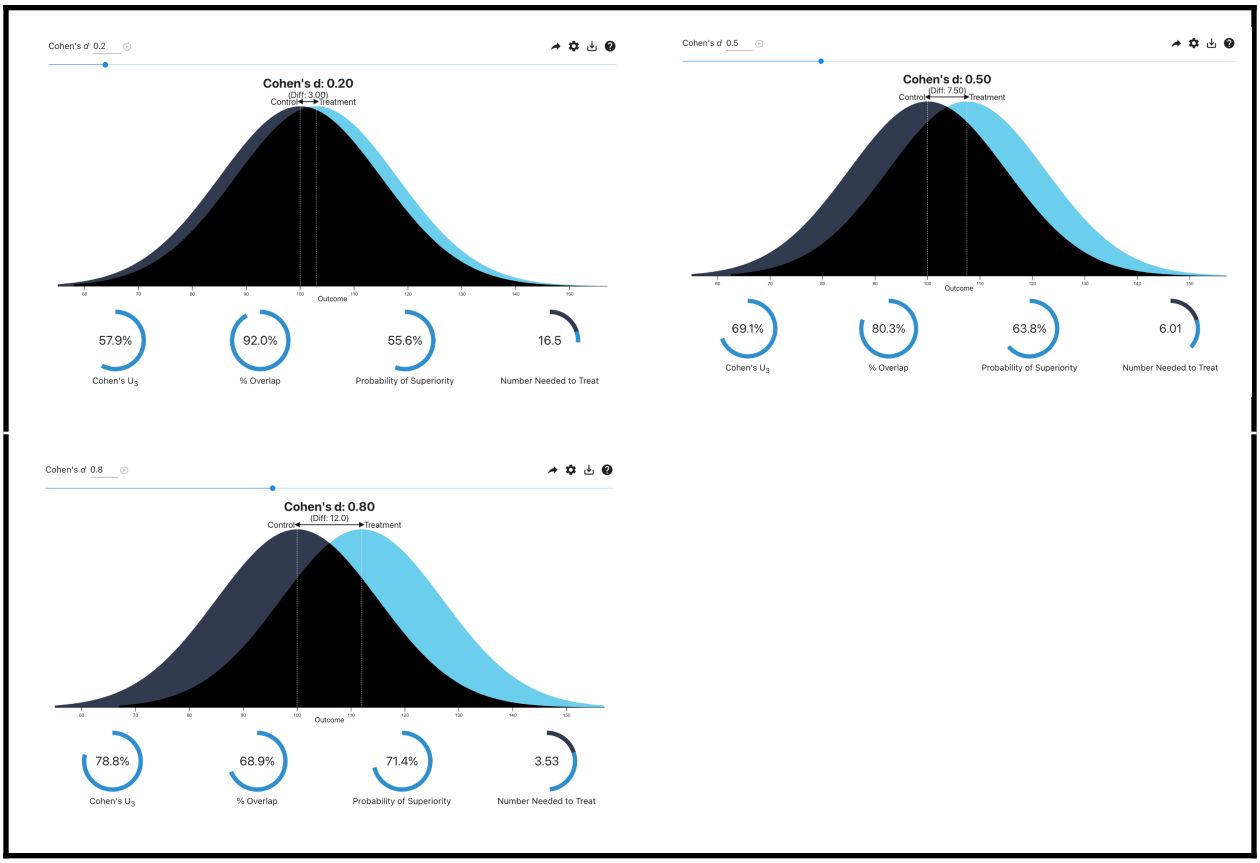
For example, imagine we wanted to test whether response times are faster after drinking caffeine. We test people after a cup of decaffeinated coffee and after a cup of regular coffee. Conducting a statistical test and looking for the *p*-value would tell us whether we can reject the null; we’re asking the question “is there a difference between these two conditions”? However, that is only half the story. There could be a significant difference between the two groups, but is the difference actually meaningful? This is where the effect size comes in. It tells us what the *actual* difference is between the two conditions.

There are two main types of effect sizes: unstandardised effects and standardised effects. Unstandardised or simple effect sizes are in the original units of measurement. In our caffeine experiment, this would be a prime example of unstandardised effects as we could calculate the mean difference between each condition. Let's say - on average - participants were 20ms faster after drinking caffeine than drinking decaf. We measured response time in milliseconds, so the effect size is just in the units of measurement. This is most useful when you have logical consistent units like milliseconds as every study using the same will be comparable.

Alternatively, we have standardised effect sizes. These convert the units of measurement into standardised units like being expressed as standard deviations. Cohen's d follows this exact process: we take the mean difference and divide it by the standard deviation of the difference. In our caffeine experiment, instead of a mean difference of 20ms, we could report the difference between conditions was d = 0.72. This is most useful when the units are inconsistent across studies or you want to compare your effect sizes to certain benchmarks. See Baguley (2009) for an overview of when you should report standardised or unstandardised effect sizes.

One such set of benchmarks are the famous cutoffs by Cohen (1988) for a small (d = 0.20), medium (d = 0.50), and large effect (d = 0.80). These benchmarks are an attractive idea as it provides a consistent standard to compare your study to. However, Cohen based them on rough estimates and they discourage critical thinking around what effect sizes are either likely or meaningful in your area of research. For example, these benchmarks underestimate effect sizes in some areas (Quintana, 2016) and overestimate them in others (Schäfer & Schwarz, 2019). This means you can use them as a starting point when you do not know any better, but ideally you would look for more topic specific benchmarks when you get to know your area of study.

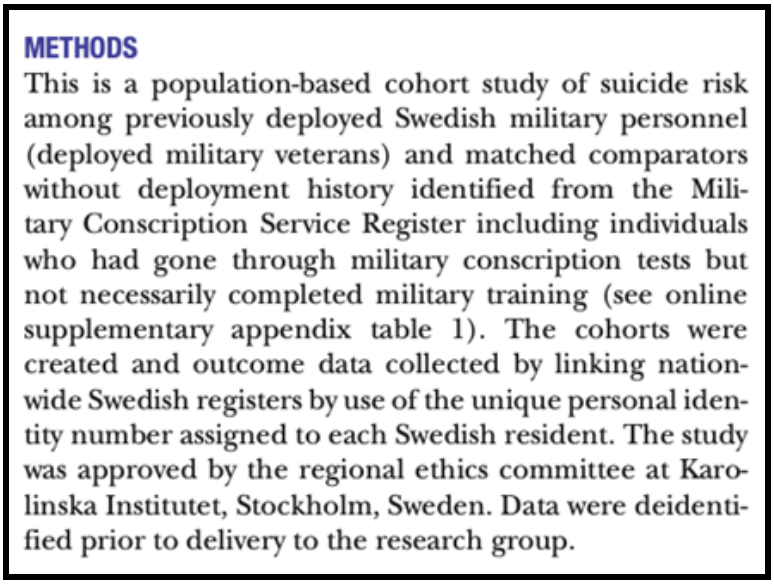
Effect sizes can feel abstract at first, so you might find this interactive visualisation helpful by Kristoffer Magnusson. The graphs below outline what a small, medium, and large effect represents if we use Cohen's guidelines. Even for large effects, there is still huge overlap between the distributions of each group.



2.4. Justifying your sample size

Before we move on to different types of power analysis, it is important to highlight this is not the only way to justify your sample size. Lakens (2021) provides an overview of different ways you can justify your sample size, with an *a priori* power analysis only being one option. Seeing sample size justification reported in articles is still disappointingly rare, so I will show some examples.

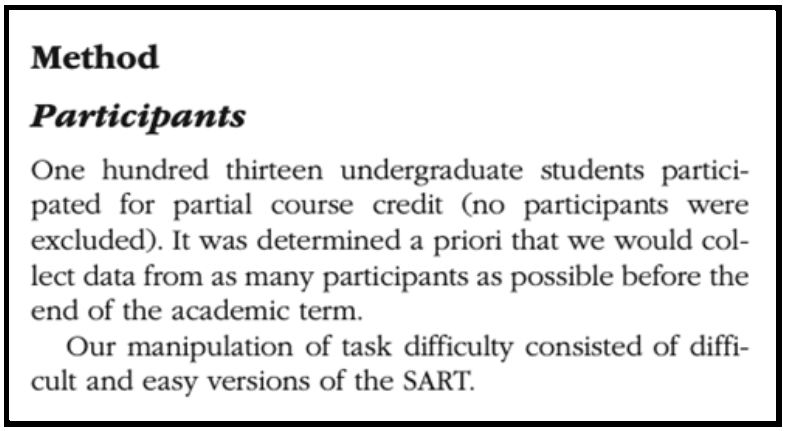
Measure the entire population: Although typically not feasible, occasionally a study can (nearly) include a whole population of interest. For example, if you have access to everyone with a rare genetic condition, your sample size is approaching the population size. Alternatively, there are cases when huge numbers of people are included. Due to conscription, Pethrus et al. (2017) had information on a whole population of Swedish people who either completed military training or not:



METHODS

This is a population-based cohort study of suicide risk among previously deployed Swedish military personnel (deployed military veterans) and matched comparators without deployment history identified from the Military Conscription Service Register including individuals who had gone through military conscription tests but not necessarily completed military training (see online supplementary appendix table 1). The cohorts were created and outcome data collected by linking nationwide Swedish registers by use of the unique personal identity number assigned to each Swedish resident. The study was approved by the regional ethics committee at Karolinska Institutet, Stockholm, Sweden. Data were deidentified prior to delivery to the research group.

Resource constraint: Many projects do not have large grant funding, so the time or money available to recruit participants is one of the deciding factors. For example, Seli et al. (2016) just recruited as many university students as they could in one academic term:



Method

Participants

One hundred thirteen undergraduate students participated for partial course credit (no participants were excluded). It was determined a priori that we would collect data from as many participants as possible before the end of the academic term.

Our manipulation of task difficulty consisted of difficult and easy versions of the SART.

Accuracy: The aim of a standard power analysis is to determine how often you would reject the null hypothesis under your parameters, but you might not simply be interested in hypothesis testing. An alternative is to focus on estimation where you want to measure an effect size with a desired level of accuracy. For example, Abt et al. (2020) outlined how to perform a power analysis to determine the width of a confidence interval around the effect size:

> Jageman, 2017). Although sample-size calculations are contextual and therefore influenced by the research design, an example using the MBESS ss.aipe.smd function is useful to highlight the approach. For a standardised mean difference (Cohen's *d*) of 0.4 between two groups, to achieve a 95% confidence interval with a width of 0.6 (0.3 either side of the point estimate) would require a sample size of at least 88. Using the median *Journal of Sports Sciences* sample size of 19 as described earlier, a confidence interval width of 1.3 (0.65 either side of the point estimate) would be achieved. This means for $d = 0.4$ the confidence interval would range from −0.25 (small negative effect) to 1.05 (large positive effect), and therefore such an interval is clearly imprecise.

***A priori* power analysis:** *A priori* power analysis is going to be our focus for the majority of this guide. If you know what effect size you are interested in, you can calculate how many participants you would need to detect it given your design. For example, in one of my own studies (Bartlett et al., 2020), I reported a simulated power analysis to justify including at least 60 participants per group:

> These values were used to conduct a simulated power analysis for a 2 x 2 mixed ANOVA using R. The code for the power analysis can be found in the pre-registration protocol (https://osf.io/am9hd/). Based on previous research, we expected non-daily smokers to display greater attentional bias towards smoking cues than daily smokers. We set the conditions of the power analysis as non-daily smokers having a 5ms (200ms) and 10ms (500ms) greater mean difference in attentional bias score than daily smokers. For each condition, the values were sampled from a normal distribution with a conservative standard deviation of 20ms. The sample size for each smoking group was increased from 10 (N = 20) to 150 (N = 300) in steps of 10, with each step repeating 10,000 times. Eighty percent power (alpha = .05, beta = .20) was reached between 50 and 60 participants per group. The target for the final sample size was 60 per group (N = 120) to avoid underestimating power.

Heuristics: Sometimes you do not have an informed decision on what sample size or effect size you choose. Historically, researchers have used heuristics such as including 20 participants per group or 50 participants plus 8 per predictor for regression. These justifications are not related to the specific design or parameters, they are just based on historical precedent. For example, Rattan et al. (2019) had a combination of heuristics and resources in their justification for at least 20 participants by the end of the university term:

> *Method*
>
> *Participants.* We aimed to recruit at least 20 participants per cell but set a strict stopping rule of the end of the school year, and therefore our sample size was determined by participant availability. A total of 53 men (4 African American/Black, 37 European American/White, 3 Latino American, 7 biracial/mixed/other, 2 unreported) were recruited from a private university in the northeastern United States.

No justification: Finally, sometimes you have no justification at all. Unless you have a time machine, you might have not considered statistical power when designing and conducting the study (maybe before reading this guide, but now you know better). You cannot go back and make up an *a priori* power analysis, so the next best thing is to think about what effect sizes your study would be sensitive to given your final sample size. Coleman et al. (2019) used this method when they did not perform a power analysis when designing the study, so the focus turned to what effects they could realistically detect with their final sample size:

> had full data for the variables included in the regression model. Given the exploratory nature of the study, an *a priori* power analysis was not conducted. However, a sensitivity power analysis (G*Power; Faul *et al.* 2009) conducted after the participants were recruited shows that 269 participants would make a linear regression model with seven predictors sensitive enough to detect a 0.03 increase in R^2 ($\alpha = .05$, power = .80).

2.5. Justifying your effect size

In the previous sections, I outlined what an effect size is and you saw you need an effect size for power analysis. The last part of this process is actually deciding what effect size you will use in your power analysis. Unfortunately, this is the single hardest part of power analysis.

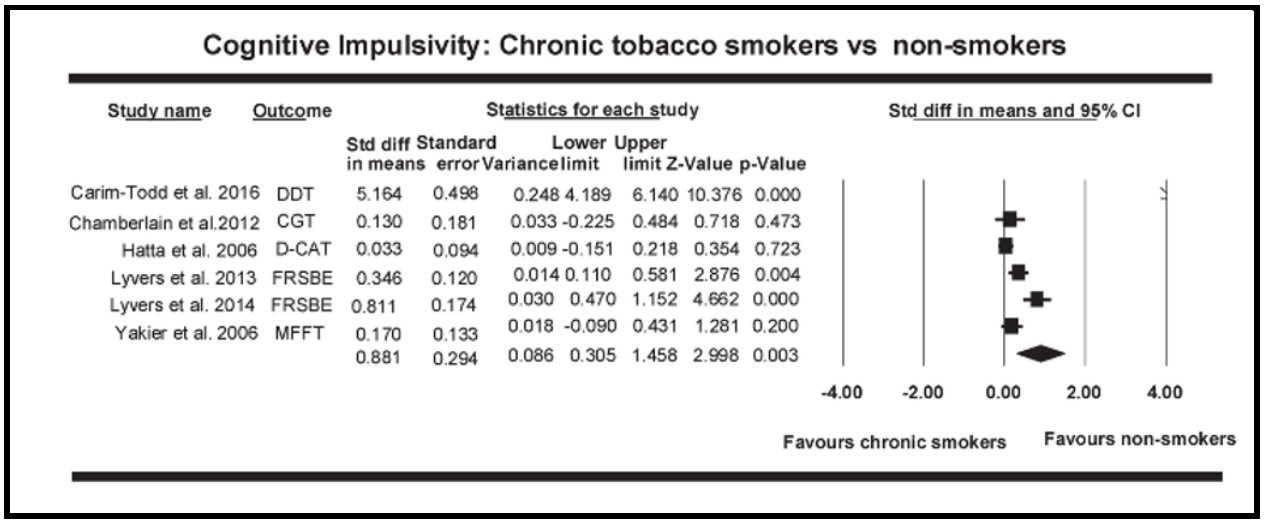
Justifying an effect size is so difficult as there is relatively little exploration in psychology research. Typically, studies focus on whether an effect is significant or not; they do not critically evaluate what effect sizes would be practically or theoretically meaningful for their area of research. When it comes to choosing an effect size for your study, Lakens (2021) outlined several strategies.

Smallest effect size of interest: One of the best options for choosing an effect size is asking yourself which effects would you not want to miss out on? Power exists along a curve (see Section 3.1.2), so if you aim for 80% power to detect an effect of d = 0.50, you would have

enough power for d = 0.50, but more power to detect larger effects and less power to detect smaller effects. Therefore, the effect size you select would be the lower bound for what you could reliably detect and any smaller effects you are implicitly saying they would be too small for you to care about. For example, in these extracts from Quintana (2018), the average effect from primary meta-analyses would have 80% power to detect d = 0.26, which researchers could use in future power analyses:

> There is no clear theoretical lower boundary for oxytocin's social-cognitive effects to determine the smallest effect size of interest (SESOI) for equivalence bounds. In lieu of this, the smallest effect size that can be detected with sufficient statistical power for a given a meta-analysis was used for the equivalence test bounds (Lakens, 2017). For this analysis, 80% power was deemed to be sufficient. For the meta-analytic effects, statistical power was calculated using a formula from Valentine and colleagues (2010). Equivalence tests using reported data were performed using the TOSTER R package (version 0.2.3; Lakens, 2017). Instead of using the default one-tailed assumption for the NHST meta-analysis, two-tailed tests were used for consistency with the original analyses.
>
> The smallest effect sizes that each meta-analysis had 80% power to detect are presented in Table 1. On average, primary meta-analyses (n = 10) had 80% power to detect an effect size of at least $d = 0.26$ (range: $d = 0.15$ to $d = 0.4$) and moderator meta-analyses (n = 18) had 80% power to detect an effect size of at least $d = 0.43$ (range: $d = 0.17$ to $d = 0.94$). Across all 28 meta-analytic tests (i.e., both primary and moderator

Expected effect size from a meta-analysis: Using the average effect size researchers reported in a meta-analysis for your area of research is one of the most popular justifications. Researchers take the effect sizes from all the studies they can find using similar methods on a topic and calculate the average effect size. This means you can use this meta-analytic effect size for your power analysis. Keep in mind how similar the studies are to your planned study, how strong the effects of publication bias are, and how consistent the effect sizes are across studies in the meta-analysis. For example, Conti et al. (2019) compared cognitive impulsivity between chronic smokers and non-smokers, and found on average non-smokers performed d = 0.88 better than chronic smokers:



Expected effect size from a key study: The average effect size from a meta-analysis is nice but what can you do if there are only a handful of studies in your area? Perhaps there is one key study you want to replicate or build on. In this scenario, you could use the effect size they reported for your power analysis. Just keep in mind how similar the study is to your design and the risk of bias if there are methodological limitations. You also have to be wary of the uncertainty around their effect size. The point estimate is just one value, and particularly with a small sample size there may be a wide confidence interval around it. Harms et al. (2018) used the effect size from a previous study as a starting point and aimed to power their experiment for an effect size twice as small (combining it with the smallest effect size of interest strategy):

> 4.2.1. Power-calculation and sample
>
> Based on the reported test statistics of the focal prime × roundedness interaction ($F_{1,314} = 13.18$) in the original article, an effect size estimate of $\eta^2 = 0.040$ was calculated. While only 250 participants were required to achieve 90% power in a 2 × 2-design with an α-level of 0.05, based on the analyses of the original findings, we assume that the true effect size is very likely to be smaller than reported (see considerations above). We thus aimed for a sample of 600 participants, which is nearly twice the original sample size of 318. This would result in a power of 90% to detect an effect $\eta^2 = 0.017$.

Expected effect size derived from a theoretical model: If there is a sufficiently developed theory in your area of research, researchers could produce a computational model to derive expected effect sizes. If it is possible, this strategy would provide strong justification for what effect size you use in your power analysis as you would be testing whether the effects are consistent with your computational model. However, I am not currently aware of this approach being used in a psychology article.

Subject specific effect size distributions: Previously, I outlined Cohen's guidelines for a small, medium, and large effect. Although these are popular, they have their limitations and

should only be used if you have nothing else to go on. An alternative is when researchers explore the effect size distributions for a specific subject area. For example, Szucs and Ioannidis (2017) found that the median small (d = 0.11), medium (d = 0.40), and large (d = 0.70) effects are smaller in cognitive neuroscience than Cohen's guidelines:

Table 1. Median and mean power to detect small, medium, and large effects in the current study and in three often-cited historical power surveys. The bottom row shows mean power computed from 25 power surveys.

		Small effect		Medium effect		Large effect	
Subfields or other surveys	**Records/Articles**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Cognitive neuroscience	7,888/1,192	0.11	0.14	0.40	0.44	0.70	0.67
Psychology	16,887/2,261	0.16	0.23	0.60	0.60	0.81	0.78
Medical	2,066/348	0.15	0.23	0.59	0.57	0.80	0.77
All subfields	26,841/3,801	0.11	0.17	0.44	0.49	0.73	0.71
Cohen (1962)	2,088/70	0.17	0.18	0.46	0.48	0.89	0.83
Sedlmeier & Gigerenzer (1989)	54 articles	0.14	0.21	0.44	0.50	0.90	0.84
Rossi (1990)	6,155/221	0.12	0.17	0.53	0.57	0.89	0.83
Rossi (1990); means of surveys	25 surveys		0.26		0.64		0.85

doi:10.1371/journal.pbio.2000797.t001

In the remainder of the guide, we will be focusing on how you can perform a power analysis to justify your sample size, either through an *a priori* power analysis when designing the study or a sensitivity power analysis when interpreting the study. You will see how you can apply these concepts to different designs, from t-tests to correlation.

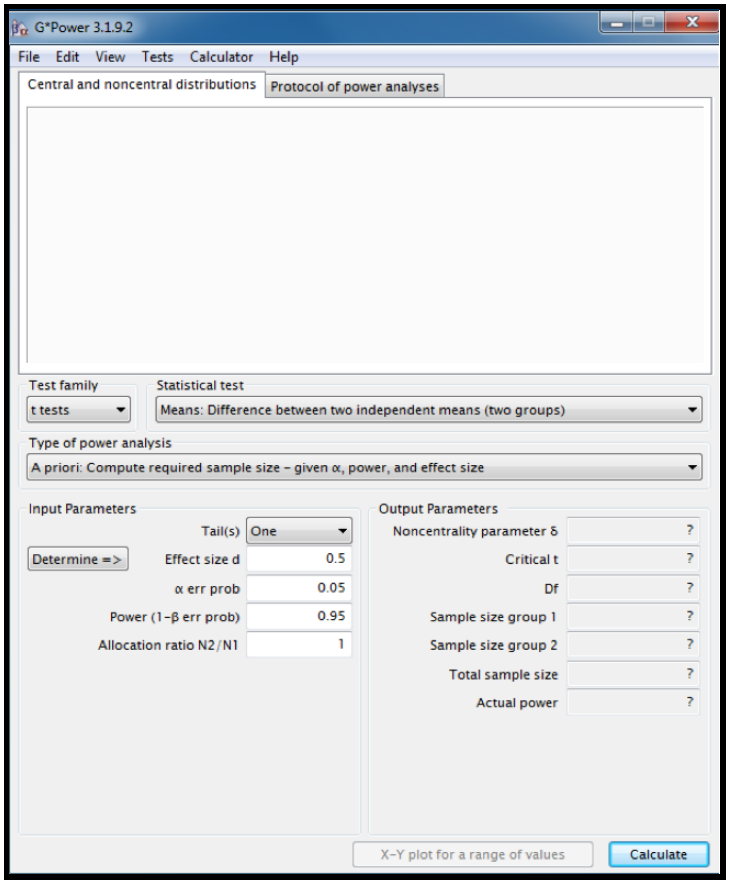
I will demonstrate them in both G*Power and jamovi in case you cannot access one of these pieces of software. I will demonstrate how you can use the Superpower Shiny app to calculate power for factorial designs. The end of the guide then outlines some supplementary information related to power analysis such as how you can calculate an effect if an article does not report one.

3. Power analysis using G*Power

3.1. T-tests

3.1.1. Independent samples t-test (*a priori*)

If you open G*Power, you should have a window that looks like this:



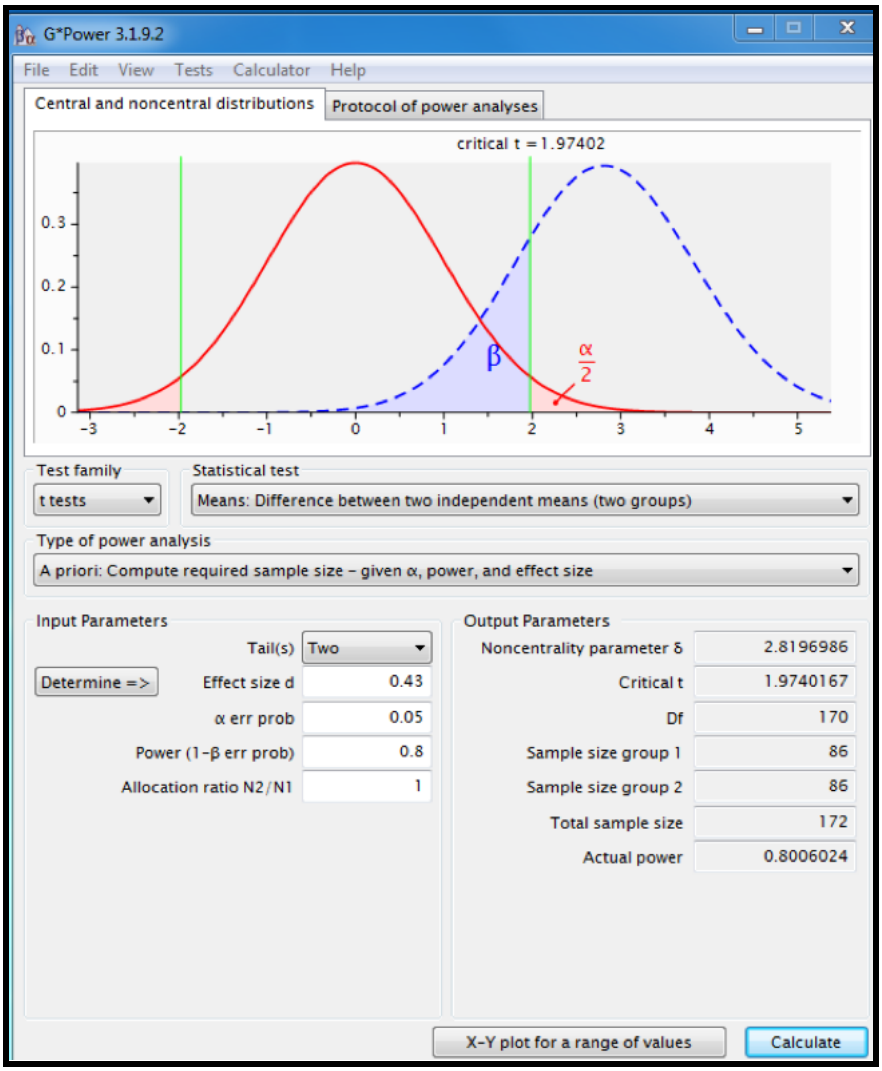
We are going to begin by seeing how you can calculate power *a priori* for an independent samples t-test. First, we will explore what each section of this window does.

- Test family - To select the family of test such as t tests, F tests (ANOVA), or χ^2. We need the default t tests for this example, so keep it as it is.
- Statistical test - To select the specific type of test. Within each family, there are several different types of test. For the t-test, you can have two groups, matched pairs, and several others. For this example, we need two groups.

- Type of power analysis - This is where we choose whether we want an *a priori* or sensitivity power analysis. For this example we want *a priori* to calculate the sample size we need in advance to detect a given effect.
- Input parameters - these are the values we need to specify to conduct the power analysis, we will go through these in turn.
    - Tails - is the test one- or two-tailed?
    - Effect size d - this is the standardised effect size known as Cohen's d. Here we can specify our smallest effect size of interest.
    - α err prob - this is our long run type one error rate which is conventionally set at .05.
    - Power (1 - β err prob) - this is our long run power. Power is normally set at .80 (80%), but some researchers argue that this should be higher at .90 (90%).
    - Allocation ratio N2 / N1 - this specifically applies to tests with two groups. If this is set to 1, sample size is calculated by specifying equal group sizes. Unequal group sizes could be specified by changing this parameter.
- Output parameters - if we have selected all the previous options and pressed calculate, this is where our required sample size will be.

The most difficult part in calculating the required sample size is deciding on an effect size (see Section 2.5). When you are less certain of the effects you are anticipating, you can use general guidelines. For example, Cohen's (1988) guidelines (e.g., small: Cohen's d = 0.2, medium: Cohen's d = 0.5, large: Cohen's d = 0.8) are still very popular. Other studies have tried estimating the kind of effects that can be expected from particular fields. For this example, we will use Richard et al. (2003) who conducted a gargantuan meta-analysis of 25,000 studies from different areas of social psychology. They wanted to quantitatively describe the last century of research and found that across all studies, the average standardised effect size was d = 0.43. We can use this as a rough guide to how many participants we would need to detect an effect of this size.

We can plug these numbers into G*Power and select the following parameters: tail(s) = two, effect size d = 0.43, α err prob = .05, Power (1 - β err prob) = 0.8, and Allocation ratio N2 / N1 = 1. You should get the following window:



This tells us that to detect the average effect size in social psychology, we would need two groups of 86 participants (N = 172) to achieve 80% power in a two-tailed test. This is probably a much bigger sample size than what you would normally find for the average t-test reported in a journal article. This would be great if you had lots of resources, but as a psychology student, you may not have the time to collect this amount of data. For modules that require you to conduct a small research project, follow the sample size guidelines in the module, but think about what sample size you would need in an ideal world based on a power analysis and reflect on the difference between the two in your discussion.

Now that we have explored how many participants we would need to detect the average effect size in social psychology, we can tinker with the parameters to see how the number of participants changes. This is why it is so important to perform a power analysis before you start collecting data, as you can explore how changing the parameters impacts the number of participants you need. This allows you to be pragmatic and save resources where possible.

- Tail(s) - if you change the number of tails to one, this decreases the number of participants in each group from 86 to 68. This saves a total of 36 participants. If your experiment takes 30 minutes, that is saving you 18 hours worth of work while still providing your experiment with sufficient power. However, using one-tailed tests can be a contentious area. See Ruxton and Neuhäuser (2010) for an overview of when you can justify using one-tailed tests.
- α err prob - setting alpha to .05 says in the long run, we want to limit the amount of type I errors we make to 5%. Some suggest this is too high, and we should use a more stringent error rate. If you change α err prob to .01, we would need 128 participants in each group, 84 more participants than our first estimate (42 more hours of data collection).
- Power (1 - β err prob) - this is where we specify the amount of type II errors we are willing to make in the long run. This also has a conventional level of .80. There are also calls for studies to be designed with a lower type II error rate by increasing power to .90. This has a similar effect to lowering alpha. If we raise Power (1 - β err prob) to .90, we would need 115 participants in each group, 58 more than our first estimate (29 more hours of data collection).

It is important to balance creating an informative experiment with the amount of resources available. This is why it is crucial that this is performed in the planning phase of a study, as these kinds of decisions can be made before any participants have been recruited.

How can this be reported?
If we were to state this in a proposal or participants section of a report, the reader needs the type of test and parameters in order to recreate your estimates. For the original example, we could report it like this:

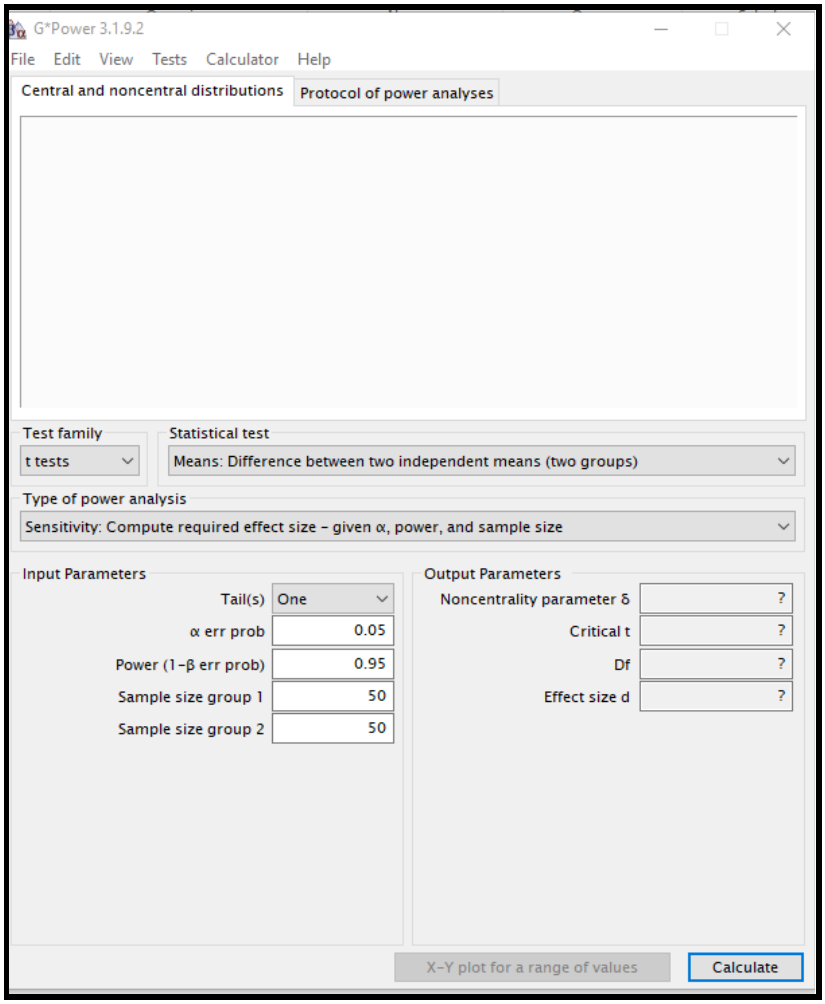
> "In order to detect an effect size of Cohen's d = 0.43 with 80% power (alpha = .05, two-tailed), G*Power suggests we would need 86 participants per group (N = 172) in an independent samples t-test". The smallest effect size of interest was set to d = 0.43 based on the meta-analysis by Richard et al. (2003).".

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement also includes your justification for the smallest effect size of interest. See Section 2.5 for different ways you can justify your choice of effect size.

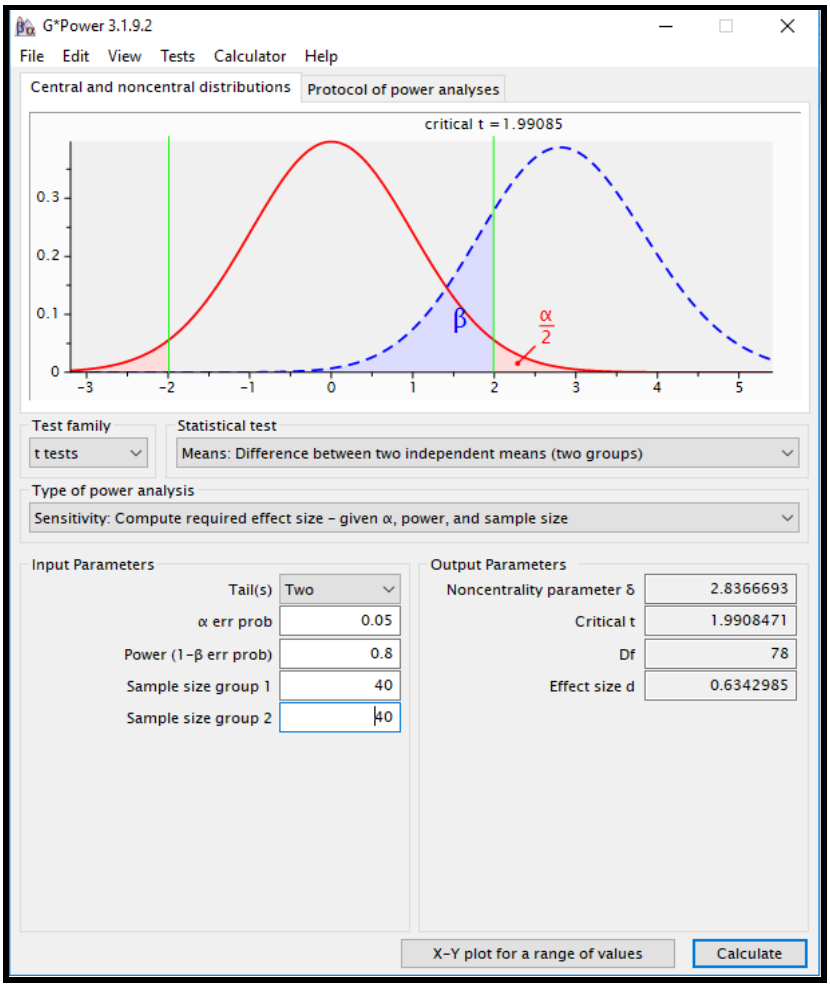
3.1.2. Independent samples t-test (sensitivity)

Selecting an effect size of interest for an *a priori* power analysis would be an effective strategy if we wanted to calculate how many participants are required before the study began. Now imagine we had already collected data and knew the sample size, or had access to a specific population of a known size. In this scenario, we would conduct a sensitivity power analysis. This

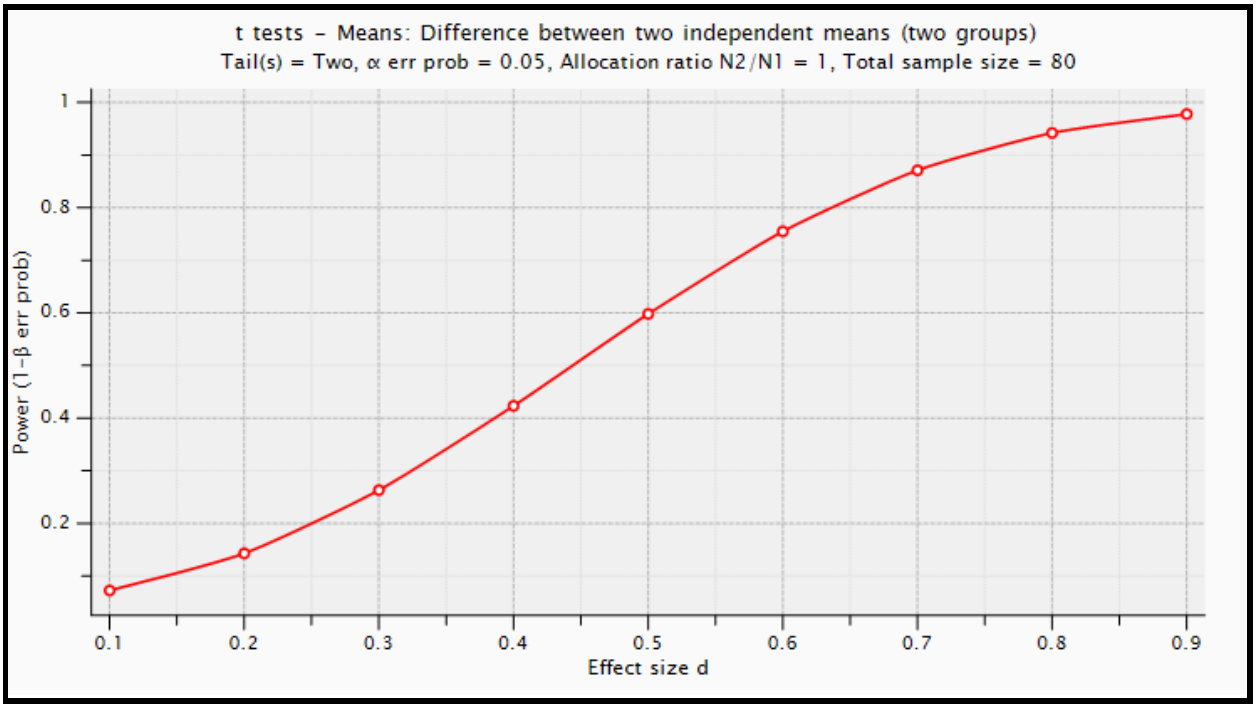
would tell us what effect sizes the study would be powered to detect in the long run for a given alpha, beta, and sample size. This is helpful for interpreting your results in the discussion, as you can outline what effect sizes your study was sensitive enough to detect, and which effects would be too small for you to reliably detect. If you change type of power analysis to sensitivity, you will get the following screen with slightly different input parameters:



All of these parameters should look familiar apart from Sample size group 1 and 2, and effect size d is now under Output Parameters. Imagine we had finished collecting data and we knew we had 40 participants in each group. If we enter 40 for both group 1 and 2, and enter the standard details for alpha (.05), power (.80), and tails (two), we get the following output:



This tells us that the study is sensitive to detect effect sizes of d = 0.63 with 80% power. This helps us to interpret the results sensibly if your result was not significant. If you did not plan with power in mind, you can see what effect sizes your study is sensitive to detect. We would not have enough power to reliably detect effects smaller than d = 0.63 with this number of participants. It is important to highlight here that power exists along a curve. We have 80% power to detect effects of d = 0.63, but we have 90% power to detect effects of approximately d = 0.73 or 50% power to detect effects of around d = 0.45. This can be seen in the following figure which you can create in G*Power using the X-Y plot for a range of values button:



This could also be done for an *a priori* power analysis, where you see the power curve for the number of participants rather than effect sizes. This is why it is so important you select your smallest effect size of interest when planning a study, as it will have greater power to detect larger effects, but power decreases if the effects are smaller than anticipated.

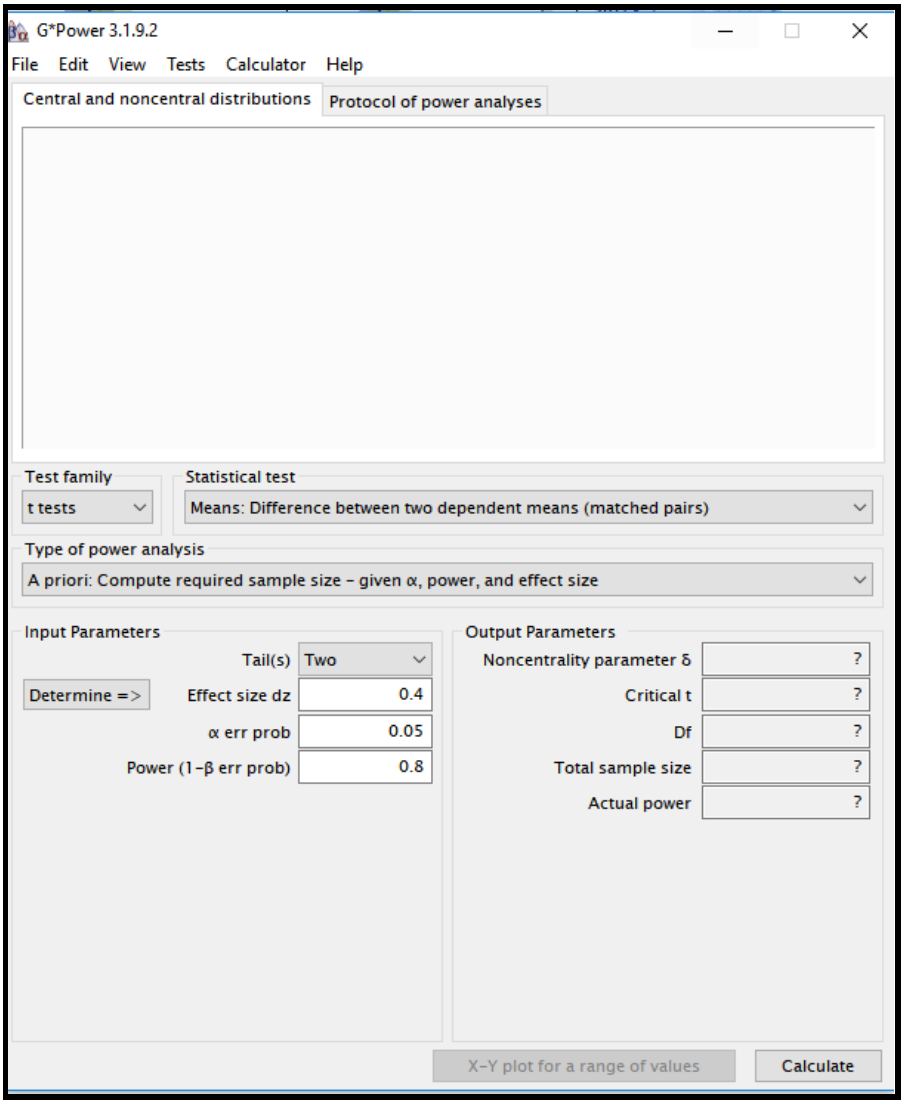
How can this be reported?
We can also state the results of a sensitivity power analysis in a report, and the best place is in the discussion as it helps you to interpret your results. For the example above, we could report it like this:

> "An independent samples t-test with 40 participants per group (N = 80) would be sensitive to effects of Cohen's d = 0.63 with 80% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.63".

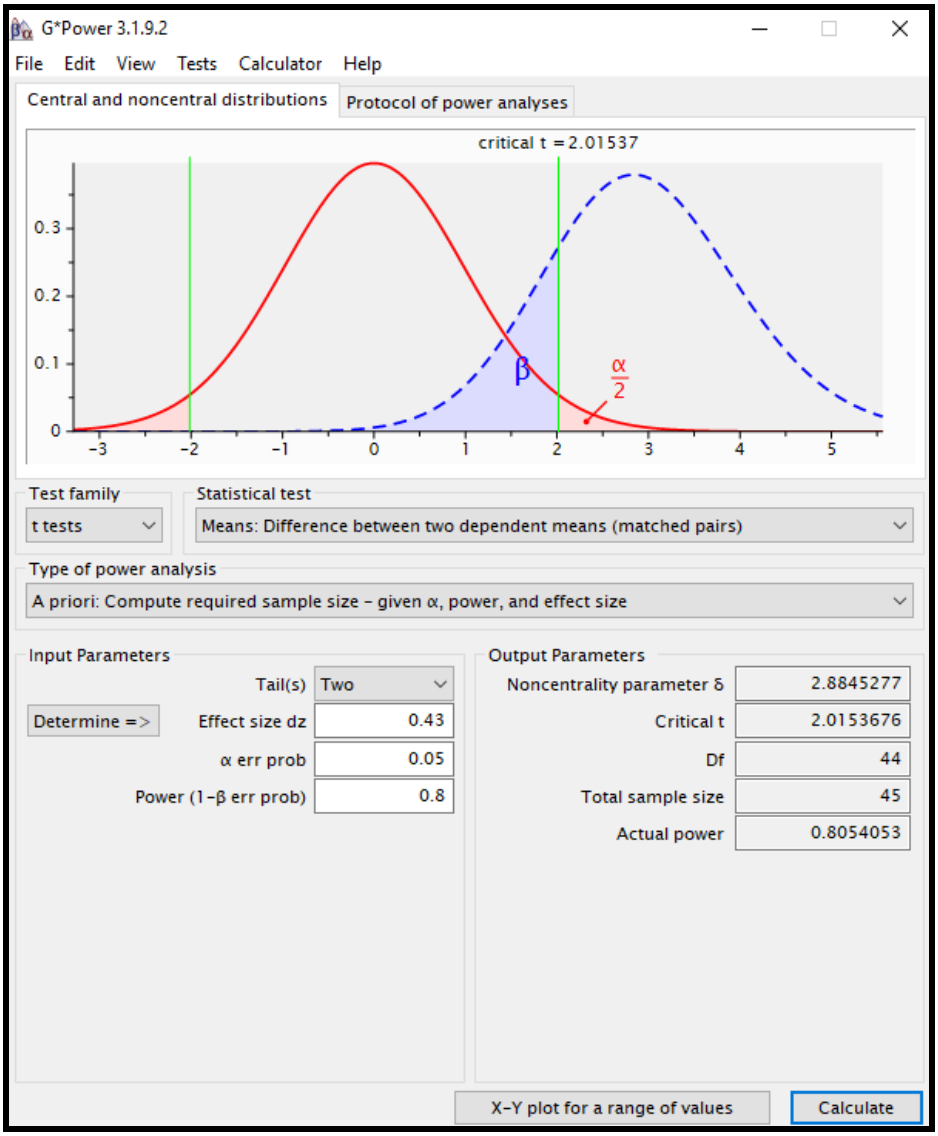
This provides the reader with all the information they would need in order to reproduce the sensitivity power analysis, and ensure you have calculated it accurately.

3.1.3. Paired samples t-test (*a priori*)

In the first example, we looked at how we could conduct a power analysis for two groups of participants. Now we will look at how you can conduct a power analysis for a within-subjects design consisting of two conditions. If you select Means (matched pairs) from the statistical test area, you should get a window like below:



Now this is even simpler than when we wanted to conduct a power analysis for an independent samples t-test. We only have four parameters as we do not need to specify the allocation ratio. As it is a paired samples t-test, every participant must contribute a value for each condition. If we repeat the parameters from before and expect an effect size of d = 0.43 (here it is called dz for the within-subjects version of Cohen's d), your window should look like this:



This suggests we would need 45 participants to achieve 80% power using a two-tailed test. This is 127 participants fewer than our first estimate (saving approximately 64 hours of data collection). This is a very important lesson. Using a within-subjects design will always save you participants for the simple reason that instead of every participant contributing one value, they are contributing two values. Therefore, it approximately halves the sample size you need to detect the same effect size (I recommend Daniël Laken's blog post to learn more). When you are designing a study, think about whether you could convert the design to within-subjects to make it more efficient.

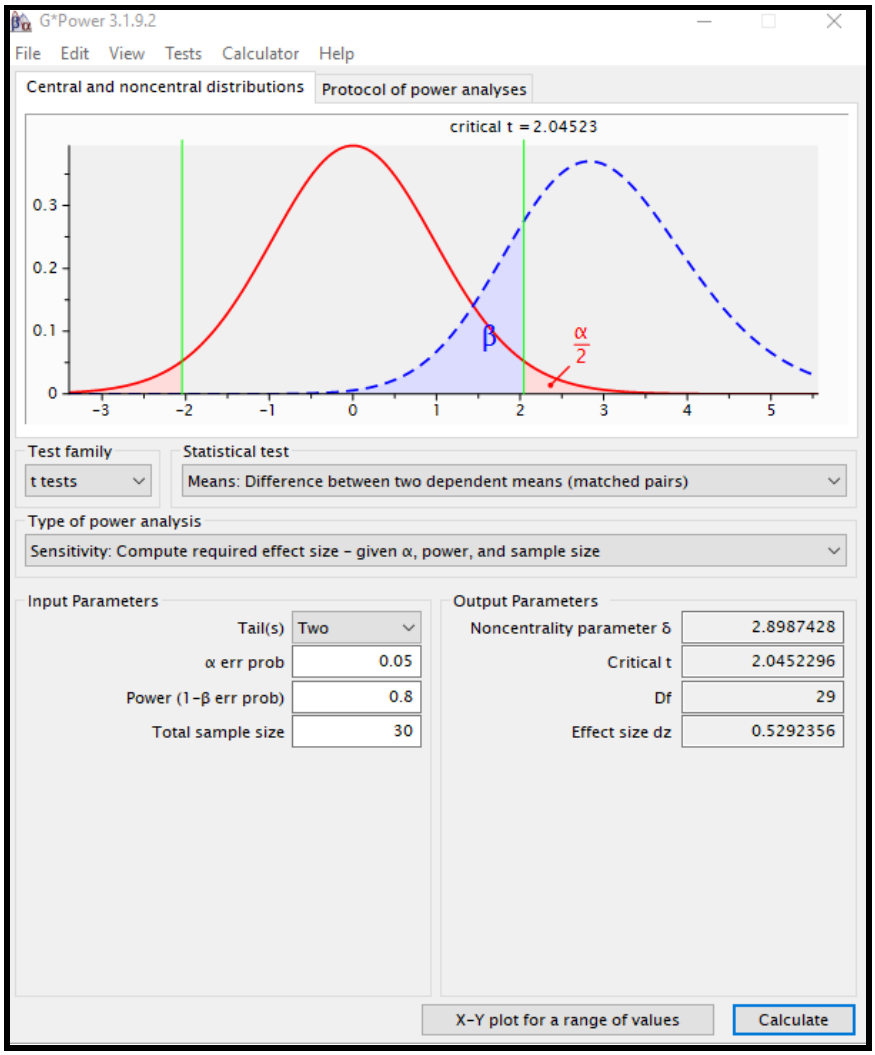
How can this be reported?
For this example, we could report it like this:

> "In order to detect an effect size of Cohen's *d* = 0.43 with 80% power (alpha = .05, two-tailed), G*Power suggests we would need 45 participants in a paired samples t-test. The smallest effect size of interest was set to d = 0.43 based on the meta-analysis by Richard et al. (2003)."

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement also includes your justification for the smallest effect size of interest. in order to reproduce the power analysis, and ensure you have calculated it accurately.

3.1.4. Paired samples t-test (sensitivity)

If we change the type of power analysis to sensitivity, we can see what effect sizes a within-subjects design is sensitive enough to detect. Imagine we sampled from 30 participants without performing an *a priori* power analysis. Set the inputs to .05 (alpha) and .80 (Power), and you should get the following output when you press calculate:



This shows that the design would be sensitive to detect an effect size of d = 0.53 with 30 participants. Remember power exists along a curve, as we would have more power for larger effects, and lower power for smaller effects. Plot the curve using X-Y plot if you are interested.

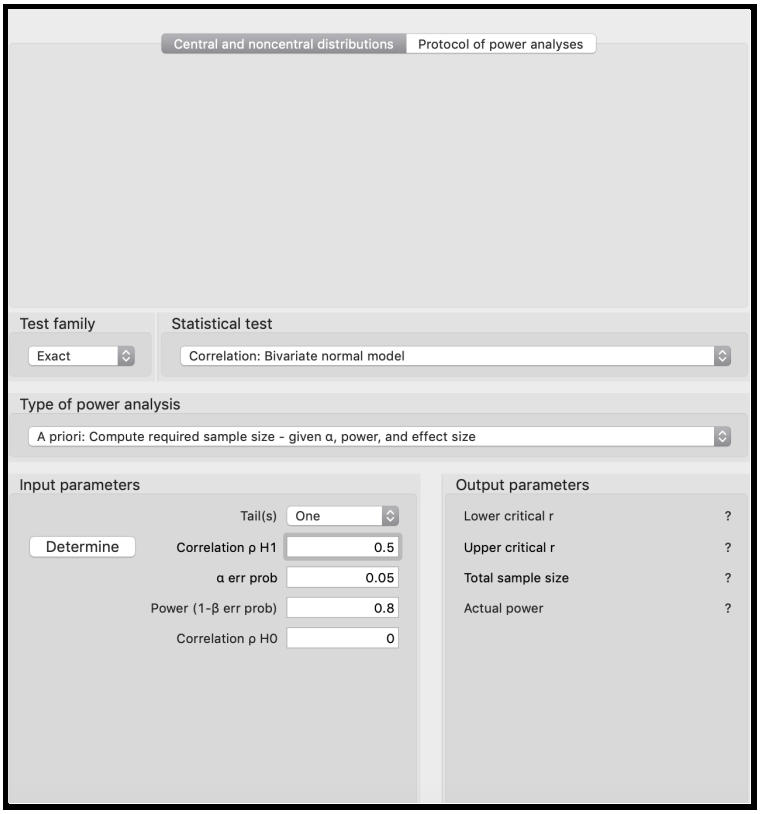
How can this be reported?
For this example, we could report it like this:

> "A paired samples t-test with 30 participants would be sensitive to effects of Cohen's d = 0.53 with 80% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.53".

3.2. Correlation

3.2.1. Correlation (*a priori*)

To work out the sample size required to detect a certain effect size, we need to select the Exact test family and correlation: bivariate normal model. You should have a window that looks like this:

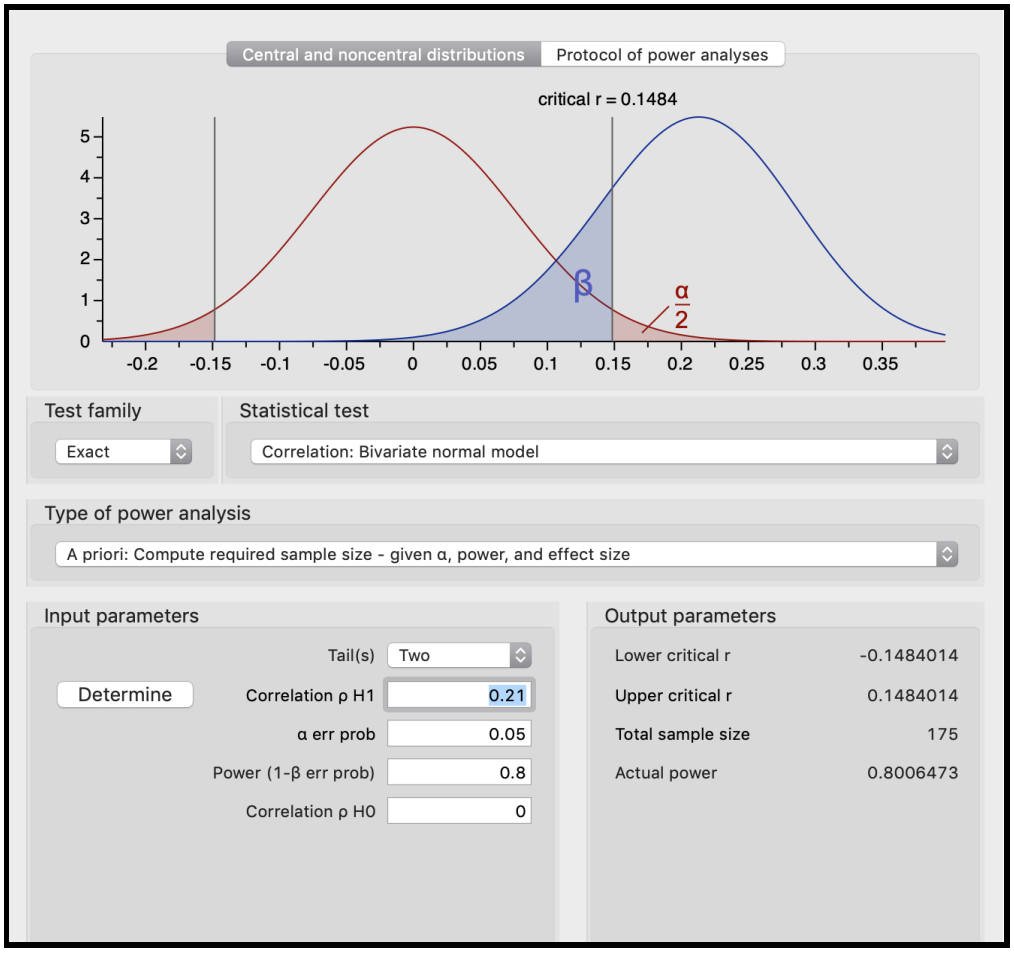


Some of the input parameters are the same as we have seen previously, but we have two new options:

- Correlation ρ H1 - This refers to the correlation you are interested in detecting. In the case of correlation, this is your smallest effect size of interest.
- Correlation ρ H0 - This refers to the null hypothesis. In most statistical software, this is assumed to be 0 as you want to test if the correlation is significantly different from 0, i.e. no correlation. However, you could change this to any value you want to compare your alternative correlation coefficient to.

For the first example, we will turn back to the meta-analysis by Richard et al. (2003). The effect size can be converted between Cohen's d and r (the correlation coefficient). If you want to convert between different effect sizes, I recommend section 13 of this online calculator.

Therefore, the average effect size in social psychology is equivalent to r = .21. If we wanted to detect a correlation equivalent to or larger than .21, we could enter the following parameters: tails (two), Correlation ρ H1 (.21), α err prob (.05), Power (0.8), and Correlation ρ H0 (0). This should produce the following window:



This suggests that we would need 175 participants to detect a correlation of .21 with 80% power. This may seem like a lot of participants, but this is what is necessary to detect a correlation this small. Similar to the t-test, we can play around with the parameters to see how it changes how many participants are required:

- Tail(s) - for a two-tailed correlation, we are interested in whether the correlation is equivalent to or larger than ±.21. However, we may have good reason to expect that the correlation is going to be positive, and it would be a better idea to use a one-tailed test. Now we would only need 138 participants to detect a correlation of .21, which would be 37 participants fewer saving 19 hours of data collection.
- Power (1 - β err prob) - Perhaps we do not want to miss out on detecting the correlation 20% of the time in the long run, and wanted to conduct the test with greater sensitivity. We would need 59 more participants (30 more hours of data collection) for a total of 234 participants to detect the correlation with 90% power (two-sided).

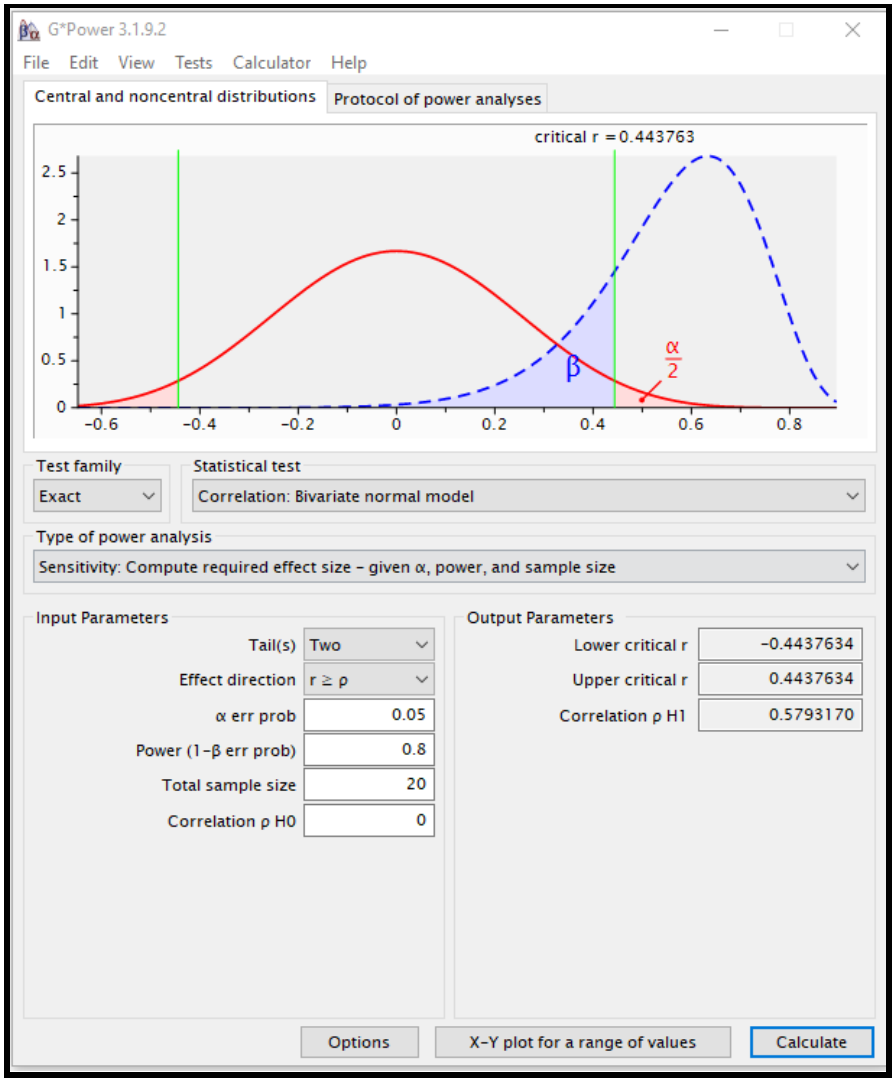
How can this be reported?
For this example, we could report it like this:

> "In order to detect a Pearson's correlation coefficient of *r* = .21 with 80% power (alpha = .05, two-tailed), G*Power suggests we would need 175 participants. The smallest effect size of interest was set to r = .21 based on the meta-analysis by Richard et al. (2003)."

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement also includes your justification for the smallest effect size of interest.

3.2.2. Correlations (sensitivity)

Like t-tests, if we know how many participants we have access to, we can see what effects our design is sensitive enough to detect. In many neuroimaging studies, researchers will look at the correlation between a demographic characteristic (e.g. age or number of cigarettes smoked per day) and the amount of activation in a region of the brain. Neuroimaging studies are typically very small as they are expensive to run, so you often find sample sizes of only 20 participants. If we specify tails (two), alpha (.05), power (.80), and sample size (20), you should get the following window:



This shows that with 20 participants, we would only have 80% power to detect correlations of r = .58 in the long run. We would only have enough power to detect a large correlation by Cohen's guidelines. Note there is a new option here called Effect direction. This does not change the size of the effect, but converts it to a positive or negative correlation depending on whether you expect it to be bigger or smaller than 0.

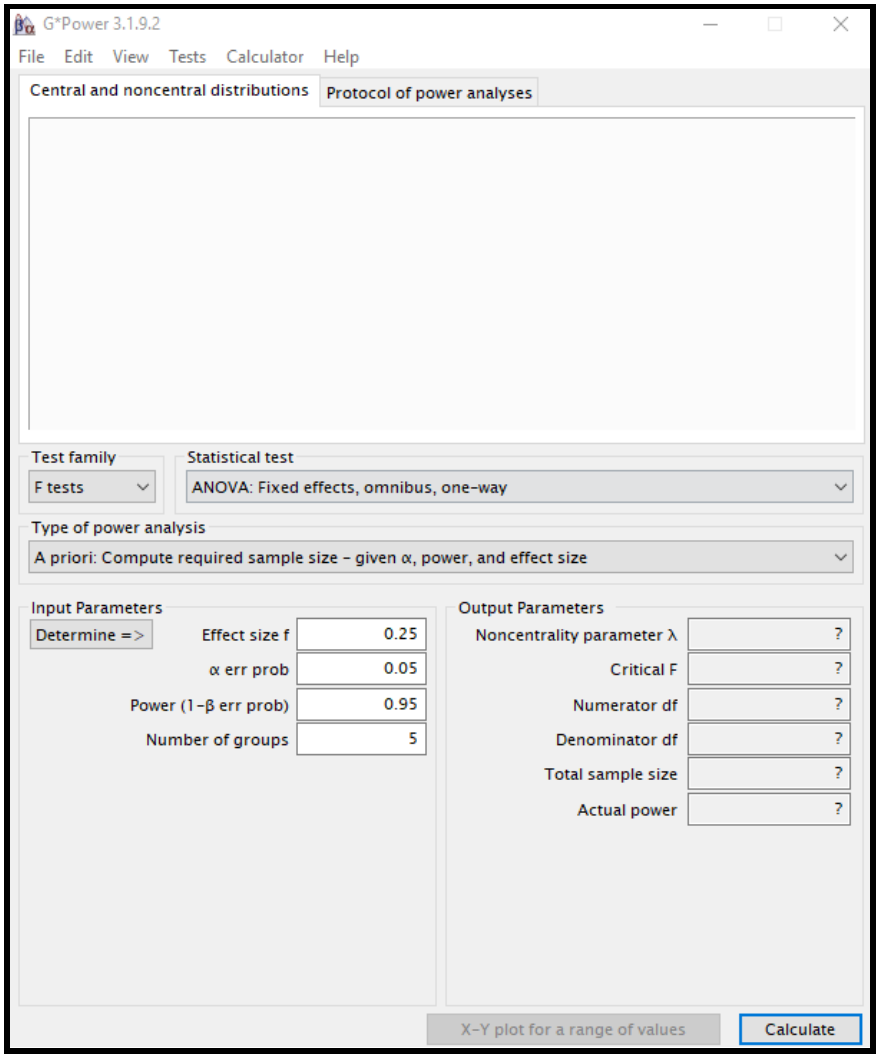
How can this be reported?
For this example, we could report it like this:

> "A Pearson's correlation coefficient with 20 participants would be sensitive to effects of *r* = .58 with 80% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect correlations smaller than *r* = .58".

3.3. Analysis of Variance (ANOVA)

3.3.1. One-way between-subjects ANOVA (*a priori*)

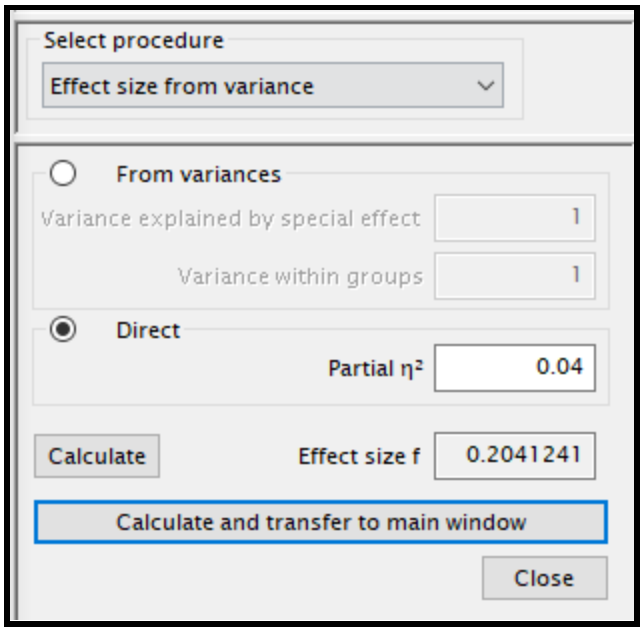
We will start with between-subjects for when we have three or more groups. In order to calculate how many participants we need, we will first need to select F tests as the Test family, and then select ANOVA: Fixed effects, omnibus, one-way. You should have a screen that looks like this:



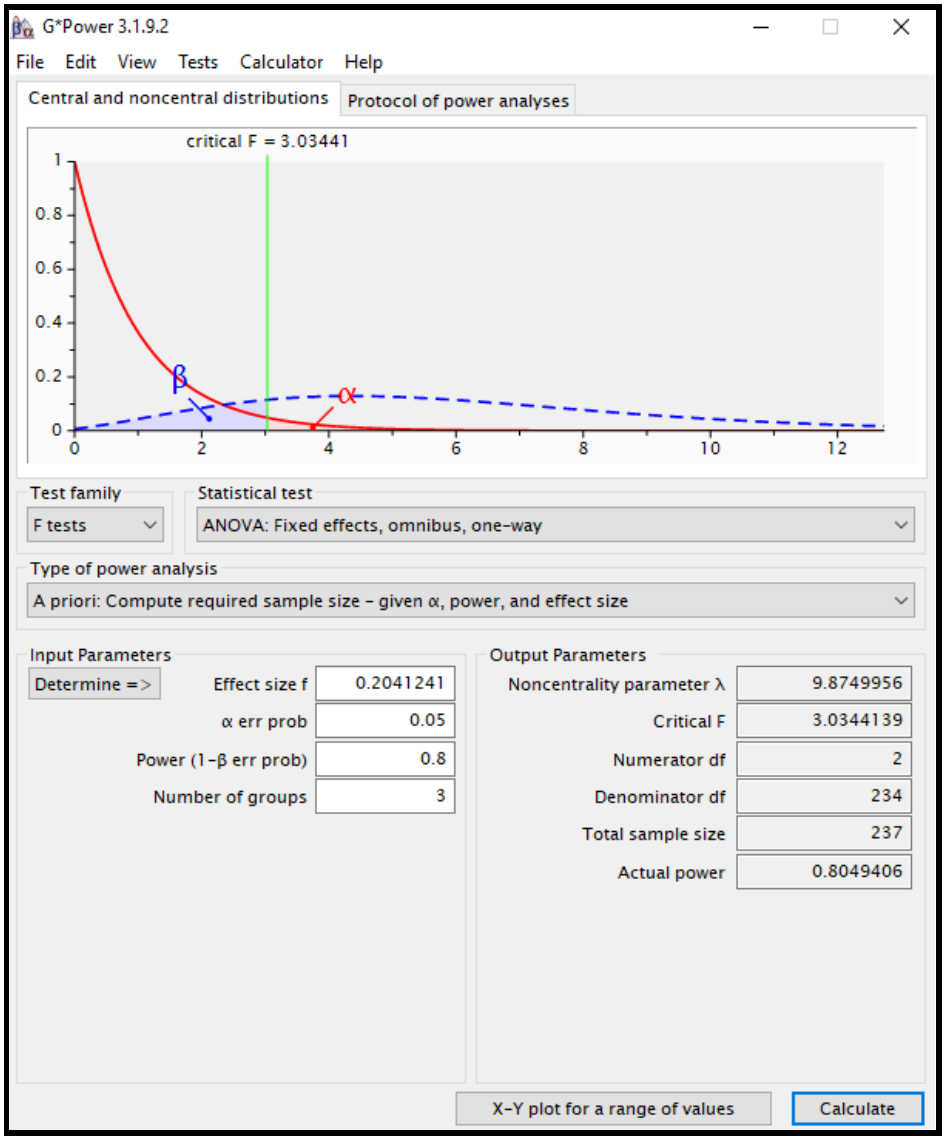
Most of the input parameters are the same as what we have dealt with for t-tests and correlation. However, we have a different effect size (Cohen's f) to think about, and we need to specify the number of groups we are interested in sampling which will normally be three or more.

ANOVA is an omnibus test that compares the means across three or more groups. This means Cohen's d would not be informative as it describes the standardised mean difference between two groups. In order to describe the average effect across many groups, there is Cohen's f. Cohen (1988) provided guidelines for this effect size too, with values of .10 (small), .25 (medium), and .40 (large). However, this effect size is not normally reported in journal articles or produced by statistics software. In its place, we normally see partial eta-squared (η^2_p) which describes the percentage of variance explained by the independent variable when the other variables are partialed out. In order words, it isolates the effect of that particular independent variable. When there is only one IV, η^2_p will provide the same result as eta-squared (η^2). Fortunately, G*Power can convert from η^2_p to Cohen's f in order to calculate the sample size.

With many effect sizes, you can convert one to the other. For example, you can convert between r and Cohen's d, and useful to us here, you can convert between Cohen's d and η^2_p. In order to convert the different effect sizes, there is section 13 of this handy online calculator. A typical effect size in psychology is η^2_p = .04 which equates to Cohen's d = 0.40. In order to use η^2_p in G*Power, we need to convert it to Cohen's f. Next to Effect size f, there is a button called Determine which will open a new tab next to the main window. From Select procedure, specify Effect size from variance, and then click Direct. Here is where you specify the η^2_p you are powering the experiment for. Enter .04, and you should have a screen that looks like this:



If you click Calculate and transfer to main window, it will input the Cohen's f value for you in the main G*Power window. Finally, input alpha (.05), power (.80), and groups (3), and you should get the following output:



This shows us that we would need 237 participants split across three groups in order to power the effect at 80%. G*Power assumes you are going to recruit equal sample sizes which would require 79 participants in each group. We can play around with some of the parameters to see how it changes how many participants are required.

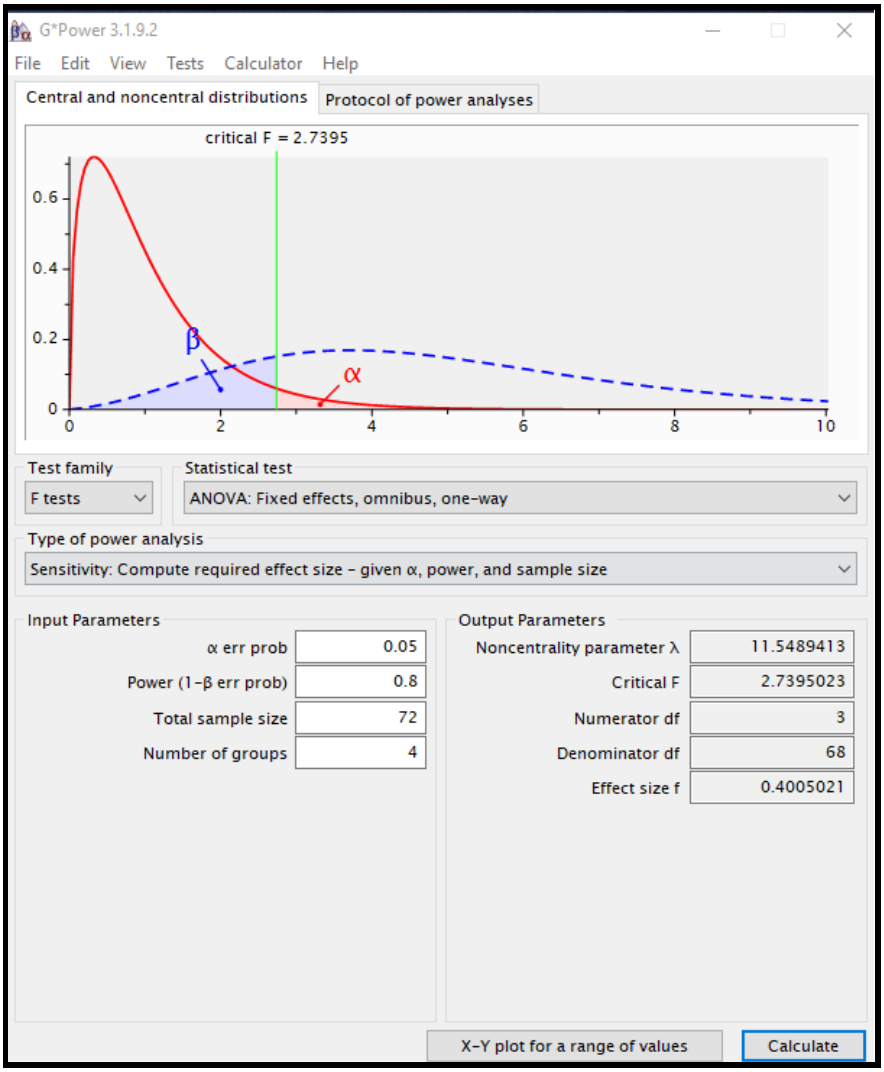
- Alpha - If we wanted to make fewer type I errors in the long-run, we could select a more stringent alpha level of .01. We would now need 339 participants (113 per group) to detect the effect with 80% power. This means 102 participants more, which would take 51 more hours of data collection.
- Power (1 - β err prob) - Perhaps we do not want to miss out on detecting the effect 20% of the time in the long run, and wanted to conduct the test with greater sensitivity. We would need 72 more participants (36 more hours of data collection) for a total of 309 participants to detect the effect with 90% power.

How can this be reported?
For this example, we could report it like this:

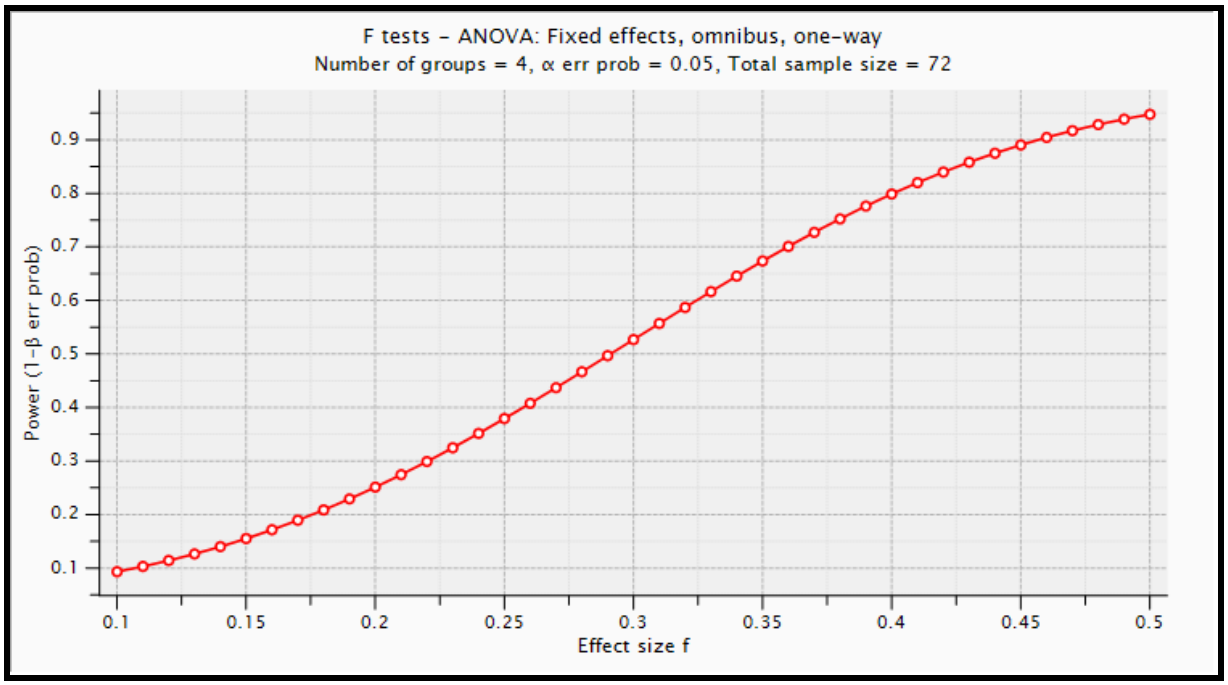
> "In order to detect an effect of η^2_p = .04 with 80% power in a one-way between-subjects ANOVA (three groups, alpha = .05), G*Power suggests we would need 79 participants in each group (N = 237)".

3.3.2. One-way between-subjects ANOVA (sensitivity)

Now that we know how many participants we would need to detect a given effect size, we can consider how sensitive a study would be if we knew the sample size. Imagine that we had 72 participants split across four groups, and we wanted to know what effect sizes this is powered to detect. Select sensitivity for type of power analysis, and enter alpha (.05), power (.80), sample size (72), and number of groups (4). You should get the following output:



This shows us that we have 80% power to detect effect sizes of Cohen's f = 0.40. This equates to a large effect, and we can convert it to η^2_p using the online calculator. This is equivalent to an effect of η^2_p = .14. As a reminder, power exists along a curve. Cohen's f = 0.40 is the smallest effect size we can detect reliably at 80% power. However, we would have greater power to detect larger effects, and lower power to detect smaller effects. It is all about what effect sizes you do not want to miss out on. The power curve for 72 participants and four groups looks like this:

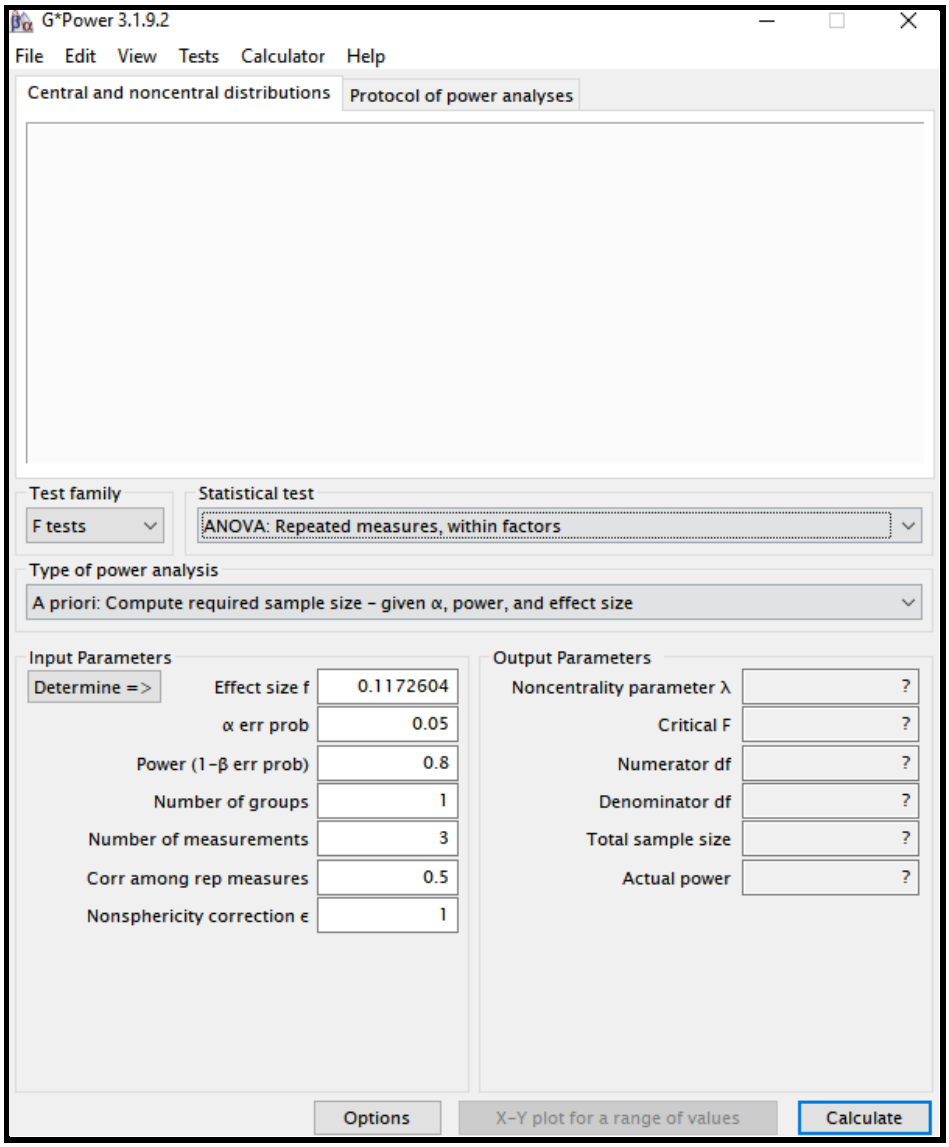


How can this be reported?
For this example, we could report it like this:

> "A one-way between-subjects ANOVA with 72 participants across four groups would be sensitive to effects of η^2_p = .14 with 80% power (alpha = .05). This means the study would not be able to reliably detect effects smaller than η^2_p = .14".

3.3.3. One-way within-subjects ANOVA (*a priori*)

Now it is time to see what we can do when we want to calculate power for three or more levels in a within-subjects design. In order to calculate power for a within-subjects design, we need to select ANOVA: Repeated measures, within factors. You should have a screen that looks like this:



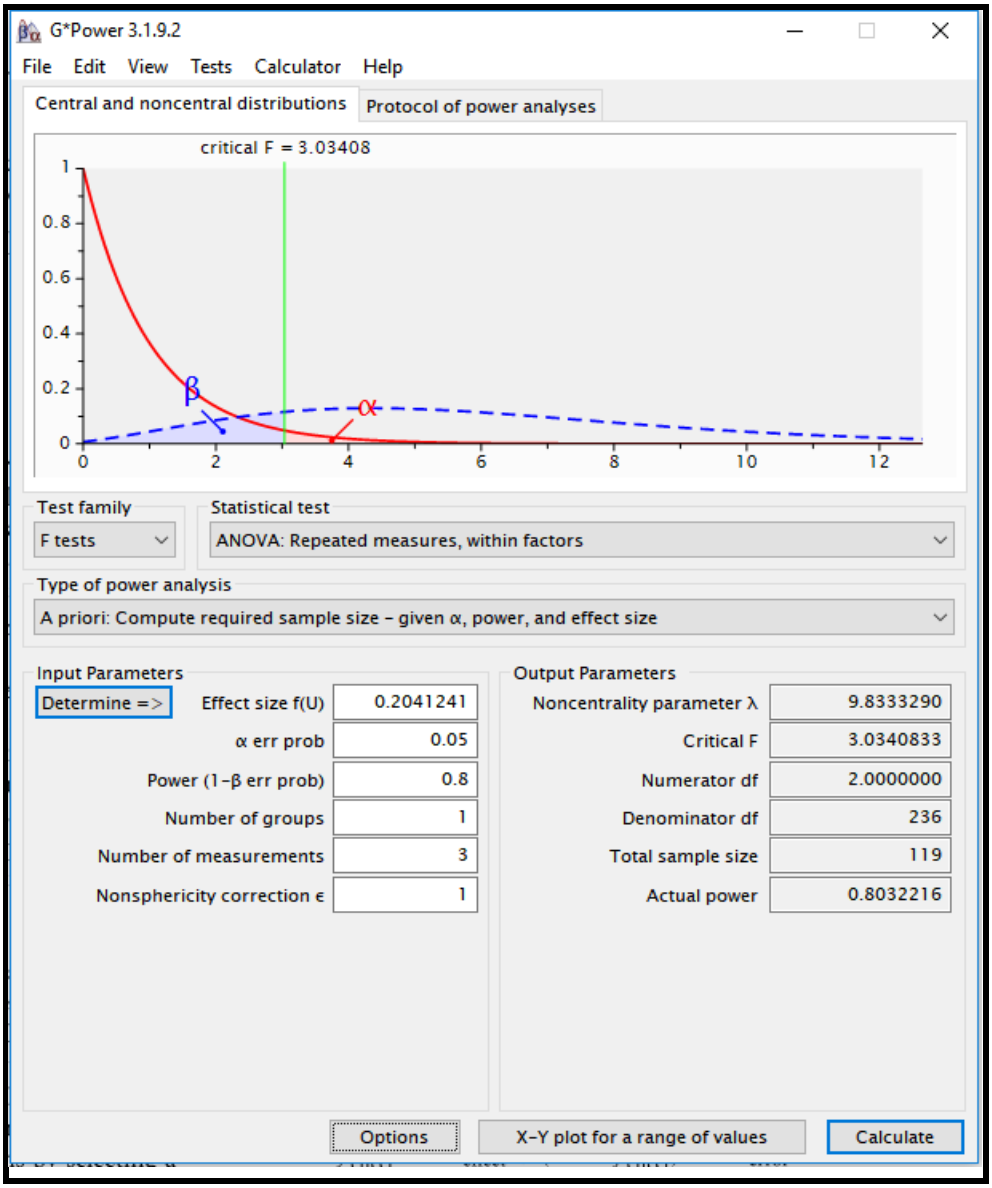
The first three input parameters should be familiar by now. The number of groups should be 1 as we have a fully within-subjects design. The number of measurements are the number of conditions we have in our within-subjects IV. To keep it simple, we will work with three conditions, so enter 3 as the number of measurements. Now we have a couple of unfamiliar parameters.

The correlation among repeated measures is something we will not need to worry about for most applications, but it's important to understand why it is here in the first place. In a within-subjects design, one of the things that affect power is how correlated the measurements are. As the measurements come from the same people on multiple conditions, they are usually correlated. If there was 0 correlation between the conditions, the sample size calculation would be very similar to a between-subjects design. As the correlation increases towards 1, the sample size you would require to detect a given effect will get smaller. The option is here as

G*Power assumes the effect size (Cohen's f) and the correlation between conditions are separate. However, if you are using η^2_p from SPSS, the correlation is already factored in to the effect size as it is based on the sum of squares. This means G*Power would provide a misleading value for the required sample size. In order to tell G*Power the correlation is already factored into the effect size, click on options at the bottom of the window and choose which effect size specification you want. For our purposes, we need as in SPSS. Select that option and click OK, and you will notice that the correlation among repeated measures parameter has disappeared. This is because we no longer need it when we use η^2_p from SPSS.

The second unfamiliar input parameter is the nonsphericity correction. If you have used a within-subjects ANOVA in SPSS, you may be familiar with the assumption of sphericity. If sphericity is violated, it can lead to a larger number of type I errors. Therefore, a nonsphericity correction (e.g. Greenhouse-Geisser) is applied to decrease the degrees of freedom which reduces power in order to control type I error rates. This means if we suspect the measures may violate the sphericity assumption, we would need to factor this into the power analysis in order to adequately power the experiment. To begin, we will leave the correction at 1 for no correction, but later we will play around with lower values in order to explore the effect of a nonsphericity correction on power.

For the first power analysis, we will use the same typical effect size found in psychology as the between-subjects example. Click determine, and enter .04 for partial η^2 (make sure effect size is set to SPSS in options). Click calculate and transfer to main window to convert it to Cohen's f. We will keep alpha (.05) and power (.80) at their conventional levels. Click calculate to get the following window:



This shows that we would need 119 participants to complete three conditions for 80% power. If we compare this to the sample size required for the same effect size in a between-subjects design, we would need 118 fewer participants than the 237 participants before. This would save 59 hours worth of data collection. This should act as a periodic reminder that within-subjects designs are more powerful than between-subjects designs.

Now it is time to play around with the parameters to see how it affects power.

- Number of measurements - One of the interesting things you will find is if we recalculate this for four conditions instead of three, we actually need *fewer* participants. We would need 90 participants to detect this effect across four conditions for 80% power. This is because each participant is contributing more measurements, so the total number of observations increases.

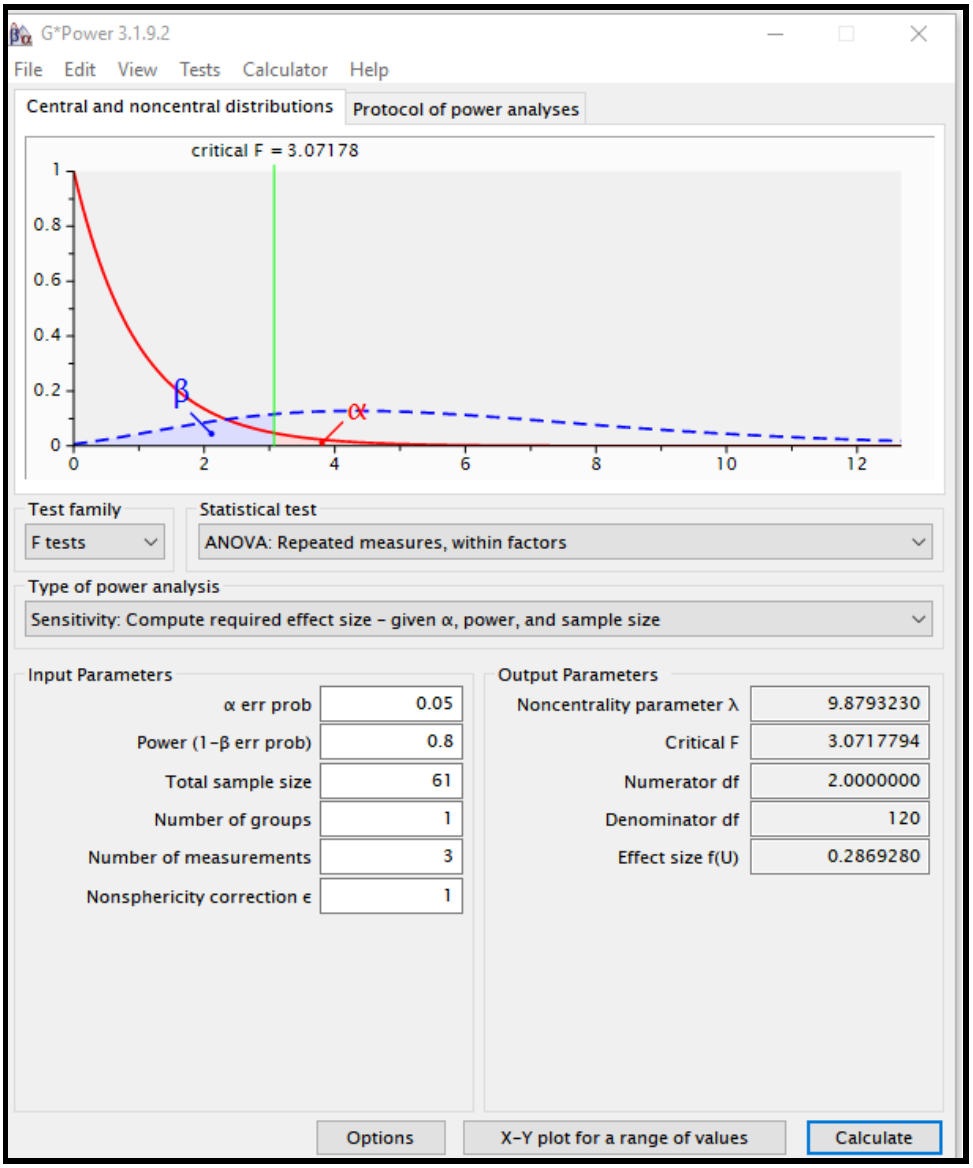
- Nonsphericity correction - Going back to three conditions, we can see the effect of a more stringent nonsphericity correction by decreasing the parameter. If we have three conditions, this can range from 0.5 to 1, with 0.5 being the most stringent correction (for a different number of conditions, the smallest lower bound can be calculated by 1 / m - 1, where m is the number of conditions. So for four conditions, it would be 1 / 3 = 0.33, but we would use 0.34 as it must be bigger than the lower bound). If we selected 0.5, we would need 192 participants to detect the effect size across three conditions. This is 73 more participants (37 more hours of data collection) than if we were to assume we do not need to correct for nonsphericity. You might be wondering how you select the value for nonsphericity correction. Hobson and Bishop (2016) have a supplementary section of their article dedicated to their power analysis. This is a helpful source for seeing how a power analysis is reported in a real study, and they choose the most stringent nonsphericity correction. This means they are less likely to commit a type II error as they may be overestimating the power they need, but this may not be feasible if you have less resources. A good strategy is exploring different values and thinking about the maximum number of participants you can recruit in the time and resources you have available.

How can this be reported?
For the first example, we could report it like this:

> “In order to detect an effect of partial eta squared = .04 with 80% power in a one-way within-subjects ANOVA (three groups, alpha = .05, non-sphericity correction = 1), G*Power suggests we would need 119 participants”.

3.3.4 One-way within-subjects ANOVA (sensitivity)

The final thing to cover for one-way ANOVA is to explore how sensitive a within-subjects design would be once we know the sample size we are dealing with. Change type of power analysis to sensitivity. If we did not conduct an *a priori* power analysis, but ended up with 61 participants and three conditions, we would want to know what effect sizes we can reliably detect. If we retain the same settings, and include 61 as the total sample size, we get the following output once we click calculate:



This shows us that we would have 80% power to detect effect sizes of Cohen's f = .29. This corresponds to η^2_p = .08 or a medium effect size.

How can this be reported?
For this example, we could report it like this:

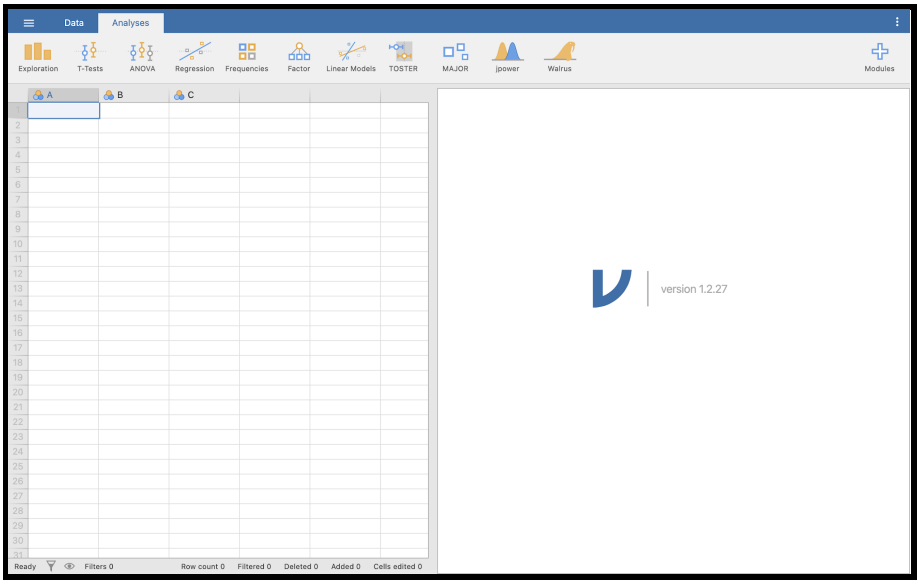
> "A one-way within-subjects ANOVA with 61 participants across three conditions would be sensitive to effects of η^2_p = .08 with 80% power (alpha = .05). This means the study would not be able to reliably detect effects smaller than η^2_p = .08".

4. Power analysis using jamovi

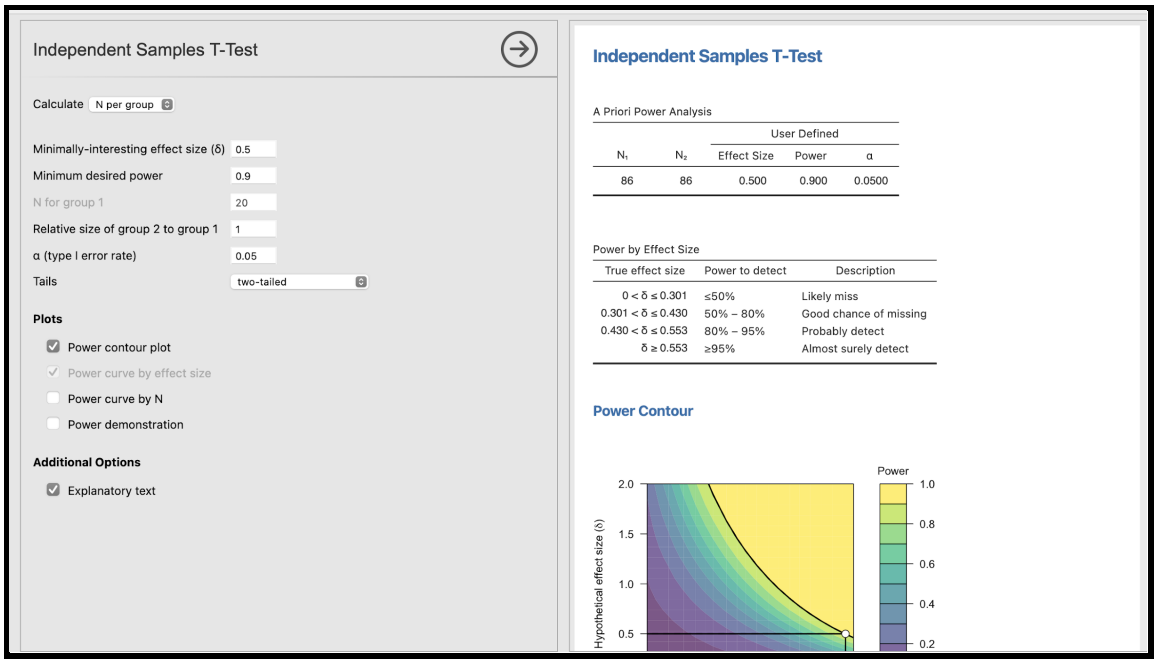
4.1. t-tests

4.1.1. Independent samples t-test (*a priori*)

If you open jamovi, you should have a window that looks like this:



If you do not have the jpower menu option, you will need to go to modules (top right) > jamovi library > install jpower. For the independent samples t-test, click on jpower and select Independent Samples T-Test. This will open the following window:



We will start by seeing how you can calculate power *a priori* for an independent samples t-test. First, the main menu options in this window are.

- Calculate - your choice of calculating N per group, power, or effect size.
- Minimally-interesting effect size - this is the standardised effect size known as Cohen's d. Here we can specify our smallest effect size of interest.
- Minimum desired power - this is our long run power. Power is normally set at .80 (80%), but some researchers argue that this should be higher at .90 (90%) or .95 (95%).
- N for group 1 - this is currently blanked out as we are calculating the sample size, but here you would define how many participants are in the first group.
- Relative size of group 2 to group 1 - If this is set to 1, sample size is calculated by specifying equal group sizes. Unequal group sizes could be specified by changing this parameter (e.g., 1.5 would mean group 2 is 1.5 times larger than group 1).
- α (type I error rate) - this is our long run type one error rate which is conventionally set at .05.
- Tails - is the test one- or two-tailed?

Like the G*Power section, we will use the meta-analytic effect size estimate of social psychology (d = 0.43) from Richard et al. (2003) as a starting point. We can use this as a rough guide to how many participants we would need to detect an effect of this size.

We can plug these numbers into jamovi and select the following parameters: effect size d = 0.43, α = .05, power = 0.8, relative size of group 2 to group 1 = 1, and two-tailed. You should get the following output:

Independent Samples T-Test

A Priori Power Analysis

		User Defined		
N_1	N_2	Effect Size	Power	α
86	86	0.430	0.800	0.0500

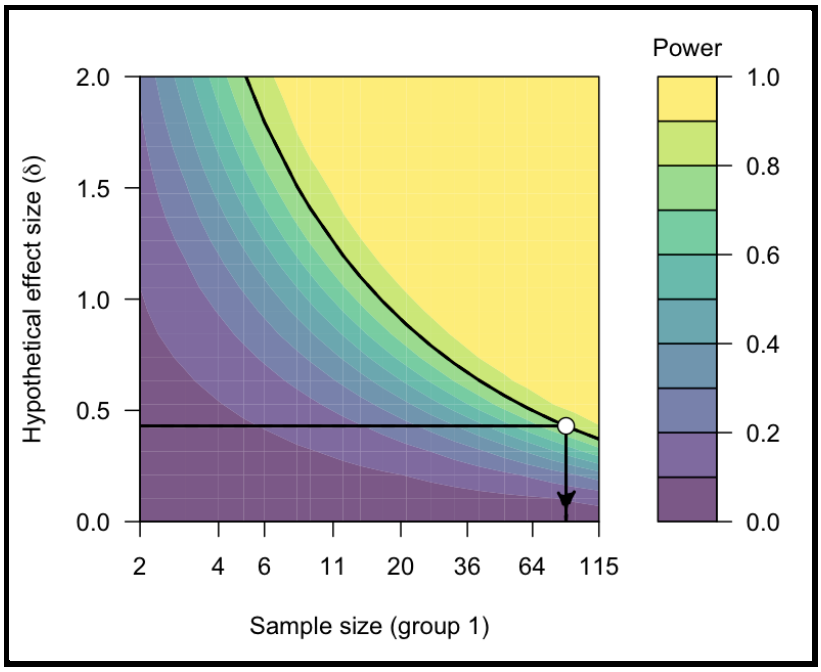
Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.301$	≤50%	Likely miss
$0.301 < \delta \leq 0.430$	50% – 80%	Good chance of missing
$0.430 < \delta \leq 0.553$	80% – 95%	Probably detect
$\delta \geq 0.553$	≥95%	Almost surely detect

This tells us that to detect the average effect size in social psychology, we would need two groups of 86 participants (N = 172) to achieve 80% power in a two-tailed test.

In contrast to G*Power, the nature of statistical power existing along a curve is much more up front. The second table in the output shows us what range of effect sizes we would likely detect with 86 participants per group. We would have 80-95% power to detect effect sizes between d = 0.43-0.55. However, we would only have 50-80% power to detect effects between d = 0.30-0.43. This shows our smallest effect of interest could be detected with 80% power, but smaller effects have lower power and larger effects would have higher power.

This is also reflected in the power contour which is reported by default. The level of power you choose is the black line that curves from the top left to the bottom right. For our specified effect size, this tells us we need 86 participants per group. For larger effects, we would have higher power and for smaller effects we would have lower power. As the sample size decreases, the power curve moves to the top left and you can detect smaller and smaller effects with your desired level of power.



Now that we have explored how many participants we would need to detect the average effect size in social psychology, we can tinker with the parameters to see how the number of participants changes. This is why it is so important to perform a power analysis before you start collecting data, as you can explore how changing the parameters impacts the number of participants you need. This allows you to be pragmatic and save resources where possible.

- Tail(s) - if you change the number of tails to one, this decreases the number of participants in each group from 86 to 68. This saves a total of 36 participants. If your experiment takes 30 minutes, that is saving you 18 hours worth of work while still providing your experiment with sufficient power. However, using one-tailed tests can be a contentious area. See Ruxton and Neuhäuser (2010) for an overview of when you can justify using one-tailed tests and ensure you preregister your prediction.
- α - setting alpha to .05 says in the long run, we want to limit the amount of type I errors we make to 5%. Some suggest this is too high and we should use a more stringent error rate. If you change α to .01, we would need 128 participants in each group (two-tailed), 84 more participants than our first estimate (42 more hours of data collection).
- Minimum desired power - this is where we specify the amount of type II errors we are willing to make in the long run. The conventional value is .80, but there are calls for studies to be designed with a lower type II error rate by increasing power to .90. This has a similar effect to lowering alpha. If we raise the minimum desired power to .90, we would need 115 participants in each group, 58 more than our first estimate (29 more hours of data collection).

It is important to balance creating an informative experiment with the amount of resources available. This is why it is crucial that this is performed in the planning phase of a study, as these kinds of decisions can be made before any participants have been recruited.

How can this be reported?
If we were to state this in a proposal or participants section of a report, the reader needs the type of test and parameters in order to recreate your estimates. For the original example, we could report it like this:

> "In order to detect an effect size of Cohen's d = 0.43 with 80% power (alpha = .05, two-tailed), the jpower module in jamovi suggests we would need 86 participants per group (N = 172) for an independent samples t-test. The smallest effect size of interest was set to d = 0.43 based on the meta-analysis by Richard et al. (2003).".

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement also includes your justification for the smallest effect size of interest. See Section 2.5 for different ways you can justify your choice of effect size.

4.1.2 Independent samples t-test (sensitivity)

Selecting an effect size of interest for an *a priori* power analysis would be an effective strategy if you wanted to calculate how many participants are required before the study began. Now imagine you had already collected data and knew the sample size, or had access to a whole population of a known size. In this scenario, we would conduct a sensitivity power analysis. This would tell us what effect sizes the study would be powered to detect in the long run for a given alpha, beta, and sample size. This is helpful for interpreting your results in the discussion, as you can outline what effect sizes your study was sensitive enough to detect and which effects would be too small for you to reliably detect. If you change Calculate to Effect size, the minimally-interesting effect size will now be greyed out as you define the sample size, alpha, and beta.

Imagine we had finished collecting data and we knew we had 40 participants in each group but did not conduct a power analysis when designing the study. If we enter 40 for N for group 1, 1 for relative size of group 2 to group 1, alpha = .05, power = .80, and tails = two, we get the following output:

A Priori Power Analysis

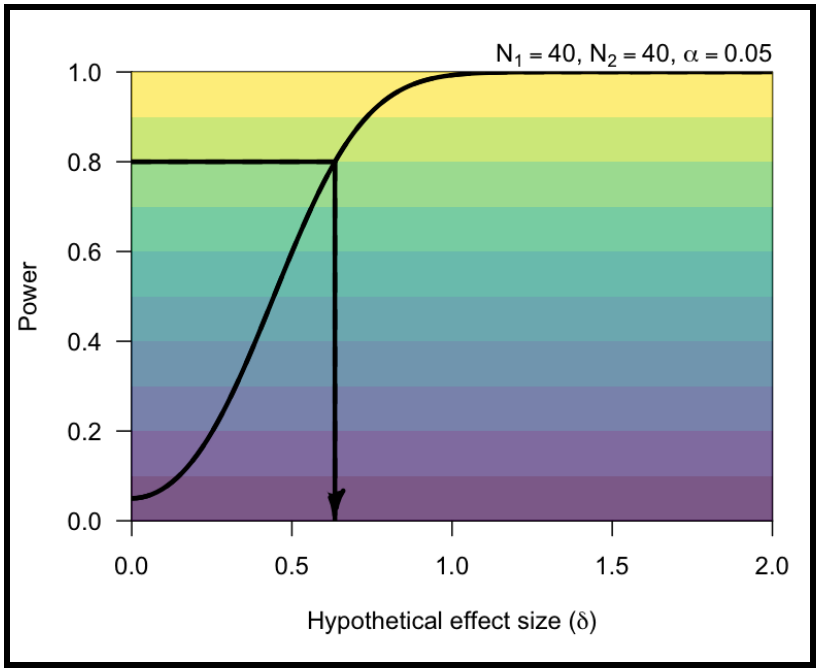
	User Defined			
Effect Size	N_1	N_2	Power	α
0.634	40	40	0.800	0.0500

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.444$	≤50%	Likely miss
$0.444 < \delta \leq 0.634$	50% – 80%	Good chance of missing
$0.634 < \delta \leq 0.816$	80% – 95%	Probably detect
$\delta \geq 0.816$	≥95%	Almost surely detect

This tells us that the study is sensitive to detect effect sizes of d = 0.63 with 80% power. This helps us to interpret the results sensibly if your result was not significant. If you did not plan with power in mind, you can see what effect sizes your study is sensitive to detect. We would not have enough power to reliably detect effects smaller than d = 0.63 with this number of participants. This is demonstrated by the power by effect size table. As the effect size gets smaller, there is less chance of detecting it with 40 participants per group.

To acknowledge how power exists along a curve, we also get a second type of graph. We now have a power curve by the effect size. This tells us how power changes as the effect size increases or decreases, with our other parameters held constant. This power curve is demonstrated below:



At 80% power, we can detect effect sizes of d = 0.63 or larger. If we follow the black curve towards the bottom left, power decreases for smaller effect sizes. This shows that once we have a fixed sample size, power exists along a curve for different effect sizes. When interpreting your results, it is important you have sufficient statistical power to detect the effects you do not want to miss out on. If the sensitivity power analysis suggests you would miss likely effects, you would need to calibrate your expectations of how informative your study is.

How can this be reported?

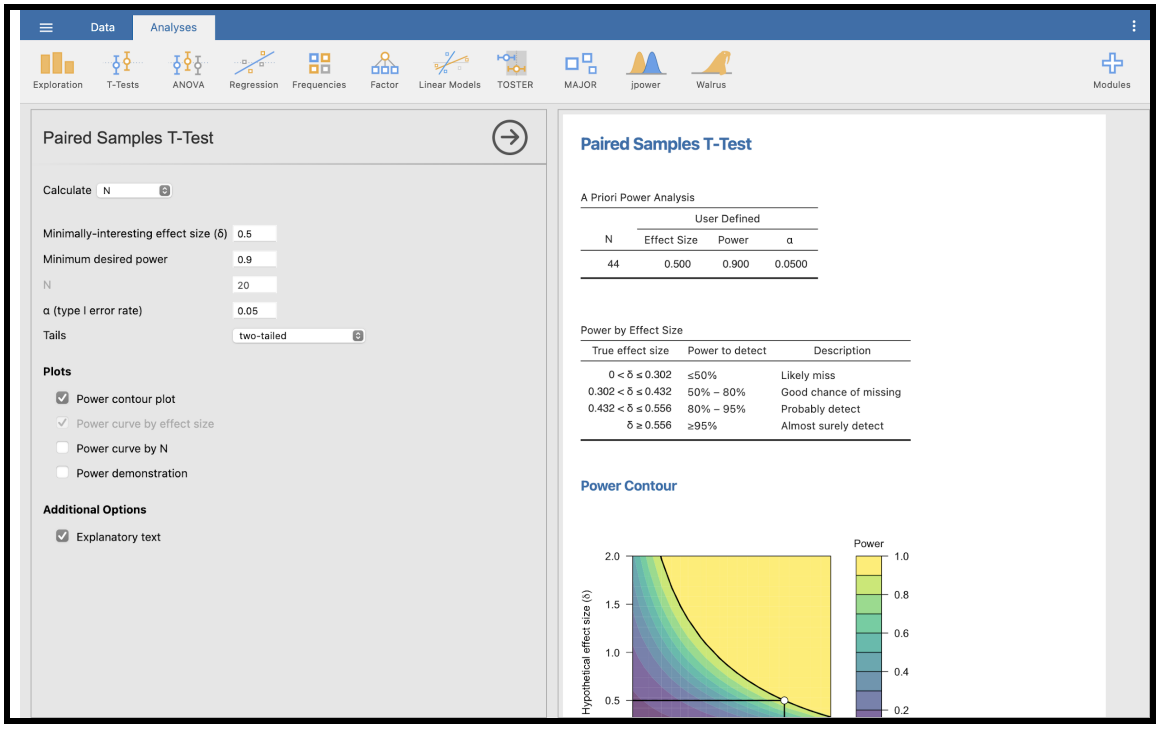
We can also state the results of a sensitivity power analysis in a report. If you did not perform an *a priori* power analysis, you could report this in the method to comment on your final sample size. If you are focusing on interpreting how informative your results are, you could explore it in the discussion. For the example above, you could report it like this:

> "The jpower module in jamovi suggests an independent samples t-test with 40 participants per group (N = 80) would be sensitive to effects of Cohen's d = 0.63 with 80% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.63".

This provides the reader with all the information they would need in order to reproduce the sensitivity power analysis and ensure you have calculated it accurately.

4.1.3. Paired samples t-test (*a priori*)

Now we will look at how you can conduct a power analysis for a within-subjects design consisting of two conditions. This time, you need to select Paired Samples T-Test from the jpower menu. You should get a window like below:



The parameters are almost identical to what we used for the independent samples t-test. We only have four parameters as we do not need to worry about the ratio of group 2 to group 1. As it is a paired samples t-test, every participant must contribute a value for each condition. If we repeat the parameters from before and expect an effect size of d = 0.43 (alpha = .05, power = .80, two-tailed), your window should look like this:

Paired Samples T-Test

A Priori Power Analysis

	User Defined		
N	Effect Size	Power	α
45	0.430	0.800	0.0500

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.299$	≤50%	Likely miss
$0.299 < \delta \leq 0.427$	50% – 80%	Good chance of missing
$0.427 < \delta \leq 0.550$	80% – 95%	Probably detect
$\delta \geq 0.550$	≥95%	Almost surely detect

This suggests we would need 45 participants to achieve 80% power using a two-tailed test. This is 127 participants fewer than our first estimate (saving approximately 64 hours of data collection).

This is a very important lesson. Using a within-subjects design will always save you participants for the simple reason that instead of every participant contributing one value, they are contributing two values. Therefore, it approximately halves the sample size you need to detect the same effect size (I recommend Daniël Laken's blog post to learn more). When you are designing a study, think about whether you could convert the design to within-subjects to make it more efficient.

How can this be reported?
For this example, we could report it like this:

> "In order to detect an effect size of Cohen's *d* = 0.43 with 80% power (alpha = .05, two-tailed), the jpower module in jamovi suggests we would need 45 participants in a paired samples t-test. The smallest effect size of interest was set to d = 0.43 based on the meta-analysis by Richard et al. (2003)".

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement should also include your justification for the smallest effect size of interest.

4.1.4. Paired samples t-test (sensitivity)

If we change Calculate to Effect size, we can see what effect sizes a within-subjects design is sensitive enough to detect. Imagine we sampled from 30 participants without performing an *a priori* power analysis. Set the inputs to alpha = .05, power = .80, and two-tailed; you should get the following output:

Paired Samples T-Test

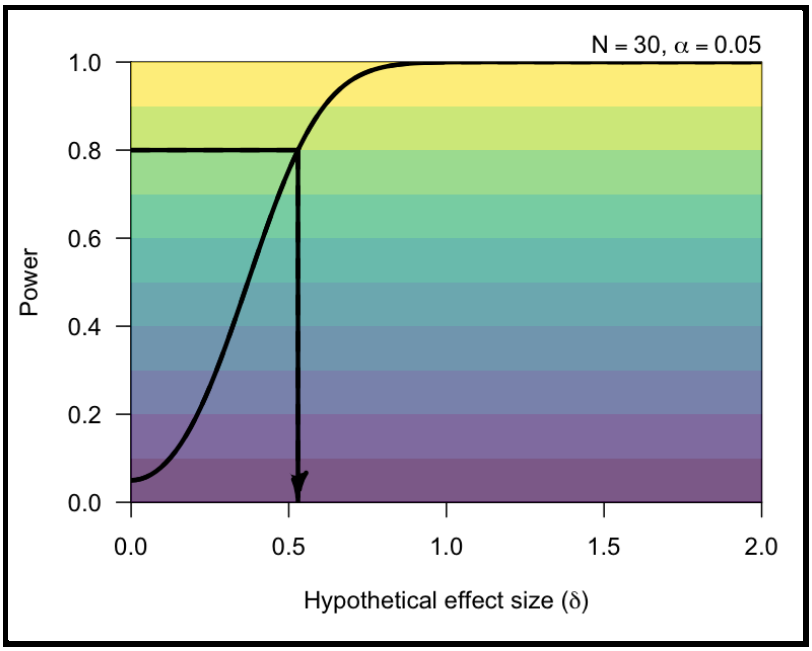
A Priori Power Analysis

	User Defined		
Effect Size	N	Power	α
0.529	30	0.800	0.0500

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.370$	≤50%	Likely miss
$0.370 < \delta \leq 0.529$	50% – 80%	Good chance of missing
$0.529 < \delta \leq 0.681$	80% – 95%	Probably detect
$\delta \geq 0.681$	≥95%	Almost surely detect

This shows that the design would be sensitive to detect an effect size of d = 0.53 with 80% power using 30 participants. Remember power exists along a curve, so we would have more power for larger effects and lower power for smaller effects:



With 30 participants, we would have 80% power to detect an effect size of d = 0.53. If we follow the black curve around, we would have lower power to detect smaller effects but higher power to detect larger effects. For an informative experiment, you should have sufficient power to detect your smallest effect size of interest.

How can this be reported?
For this example, we could report it like this:

> "The jpower module in jamovi suggests a paired samples t-test with 30 participants would be sensitive to effects of Cohen's d = 0.53 with 80% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.53".

5. Power for factorial designs

We covered one-way ANOVA in G*Power, but calculating power for factorial designs is not accurate in G*Power. There is a great blog post by Roger Giner-Sorolla for why G*Power drastically underestimates the sample you would need to power an interaction. In order to adequately power interaction effects, you ideally need to simulate your experiment. This means you would program the effect sizes you are expecting across multiple factors and see how many times it would return a significant result if you repeated it many times. This cannot be done in G*Power, so it would require you to learn a programming language like R or Python.

Fortunately, two researchers have made this process more user friendly by creating an online app to calculate power for main and interaction effects in factorial designs. There is an article explaining their app (Lakens & Caldwell, 2021) and the online app itself. This app follows an approximation approach to make it faster to calculate where you can only specify a single standard deviation, correlation (for within-subject / mixed designs), and group sample size. If you want to vary these parameters, there is a sister app available which will take longer to calculate. Caldwell et al. (2021) are currently writing a book describing the Superpower app and R package, so that will provide a more comprehensive overview of all the options.

In this next section, I will provide an overview of why power is more difficult to calculate for factorial designs. I will then demonstrate how you can use the Shiny app to calculate power for different factorial designs.

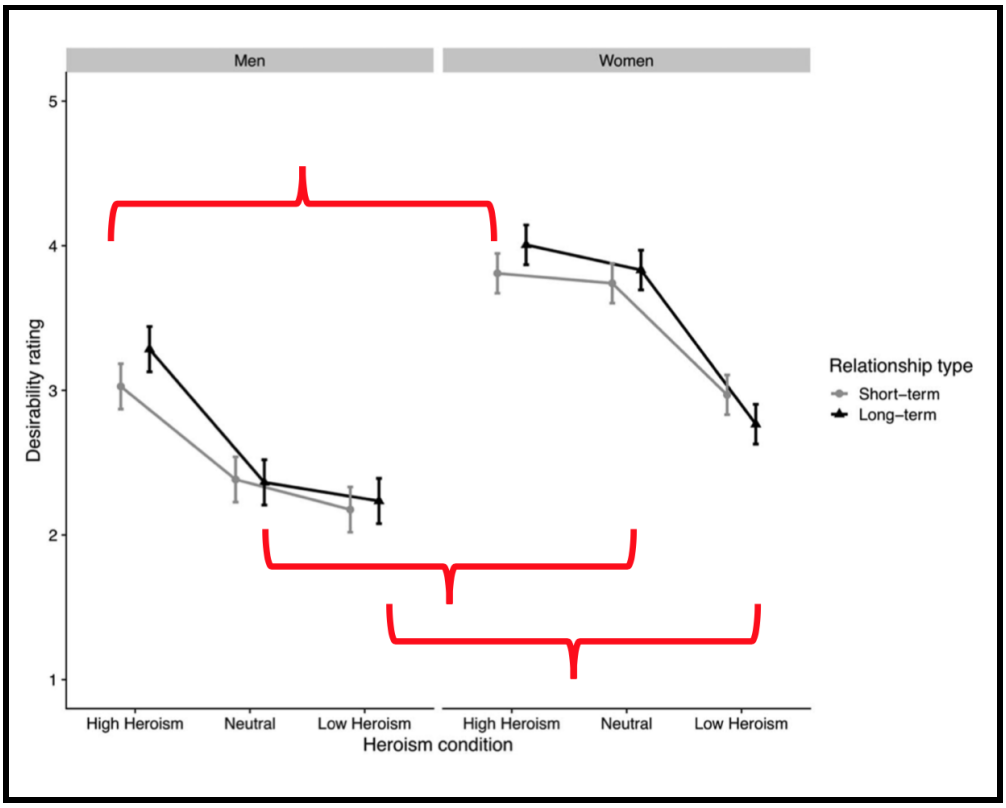
5.1. Principles of power for factorial designs

When we focus on two groups or two conditions for t-tests, power analysis is simple. You are just making one comparison: how big is the difference expected to be between these two things? This makes it difficult to apply the same approach to factorial designs as you are no longer just comparing two groups/conditions.

In the simplest example for a 2x2 design, there are three effects to compare: the main effect of IV1, the main effect of IV2, and the interaction between IV1 and IV2. There are three comparisons to make here and it is unlikely you would expect the same effect size for all three comparisons, meaning you must ensure all three effects are sufficiently powered. You might expect the difference for IV1 to be a lot larger than the difference for IV2. This means you would have to ensure the *smallest* effect is sufficiently powered. If the smallest effect is covered - remember power exists along a curve - then the larger effects would have higher power.

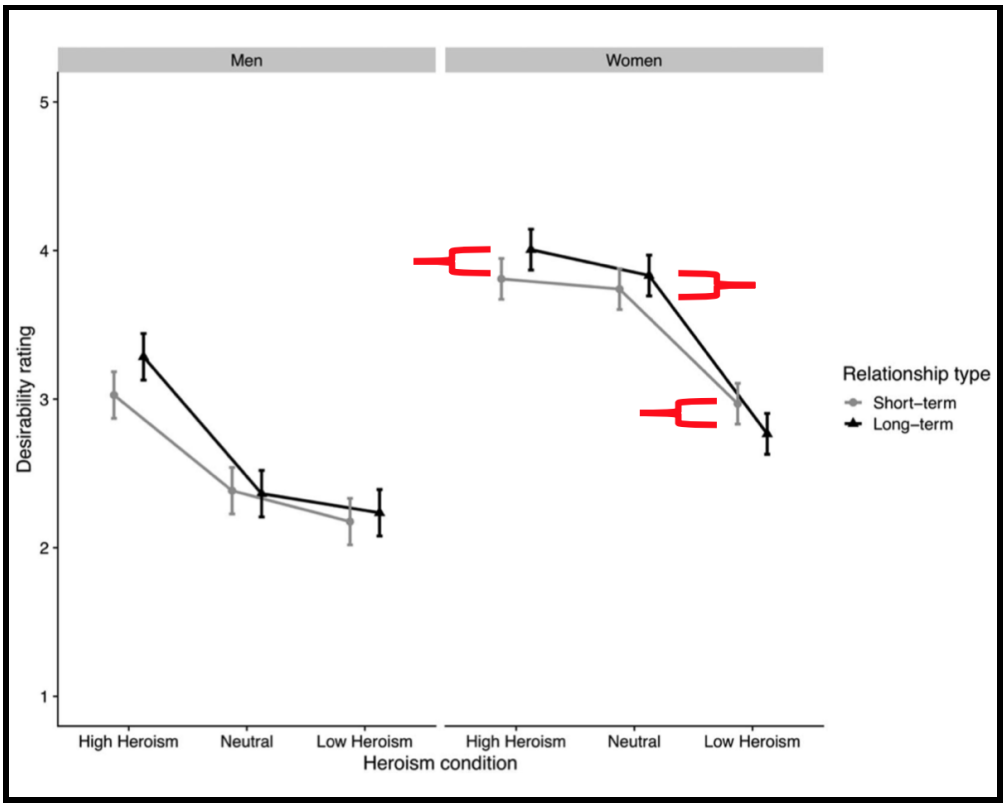
For example, consider this 2x2x3 mixed design (adapted from Bhogal & Bartlett, 2020). There is one between-subjects effect of participant gender, one within-subjects effect of heroism condition, and one within-subjects effect of relationship type. With a design this complicated, it is

unlikely you would be interested in every comparison. One of our hypotheses predicted women would report higher desirability for high heroic targets than men. This would be an interaction between gender and heroism. This means for a comparison between genders, we would be focusing on these differences:



We focused on comparing men and women for different levels of the heroism condition. We could have expected one of these comparisons to be larger than another. This means we should have powered the study for the *smallest* of these expected differences, ensuring the other larger anticipated effects have higher levels of power (for full disclosure, we actually fell into the trap of calculating power for the interaction using G*Power, but I did not know better at the time).

Conversely, one of our other predictions focused on the interaction between heroism and relationship type, expecting higher desirability ratings for a long-term relationship (compared to short-term) for high heroism. This would be like comparing the following conditions:



As this effect is fully within-subjects, it is probably going to be covered by the sample size required for the comparisons involving the between-subject IV. It is still important to consider though, as you might expect these differences to be much smaller.

When conducting a power analysis for factorial designs, it is important to keep in mind all your comparisons of interest should have sufficient power. Consider what your *smallest* effect size of interest would be and ensure the final sample reflects the most conservative estimate. It is better to overestimate power for larger effects than just focus on the best case scenario and miss out on smaller effects in your design.

Now that we have covered the principles of power analysis for factorial designs, we can see how to use the Superpower Shiny app for different designs.

5.2. Factorial between-subjects designs

For this example, I will pretend we are going to replicate experiment two of Schmeck et al. (2014). In contrast to G*Power and jamovi where you can use standardised effect sizes, factorial designs require the means and standard deviations per group. Although, there is a trick where you can define the standard deviation as 1 and the mean differences are essentially expressed as Cohen's d. This only works for fully between-subjects designs though.

Schmeck et al. (2014) investigated whether drawings can enhance the learning process. School age children learnt about how the influenza virus replicates and they could draw the process as they read about the topic with or without an author-generated visual prompt. This meant the researchers used a 2x2 between-subjects design with one IV of learner-generated drawing (yes vs no) and one IV of author-generated drawing (yes vs no).

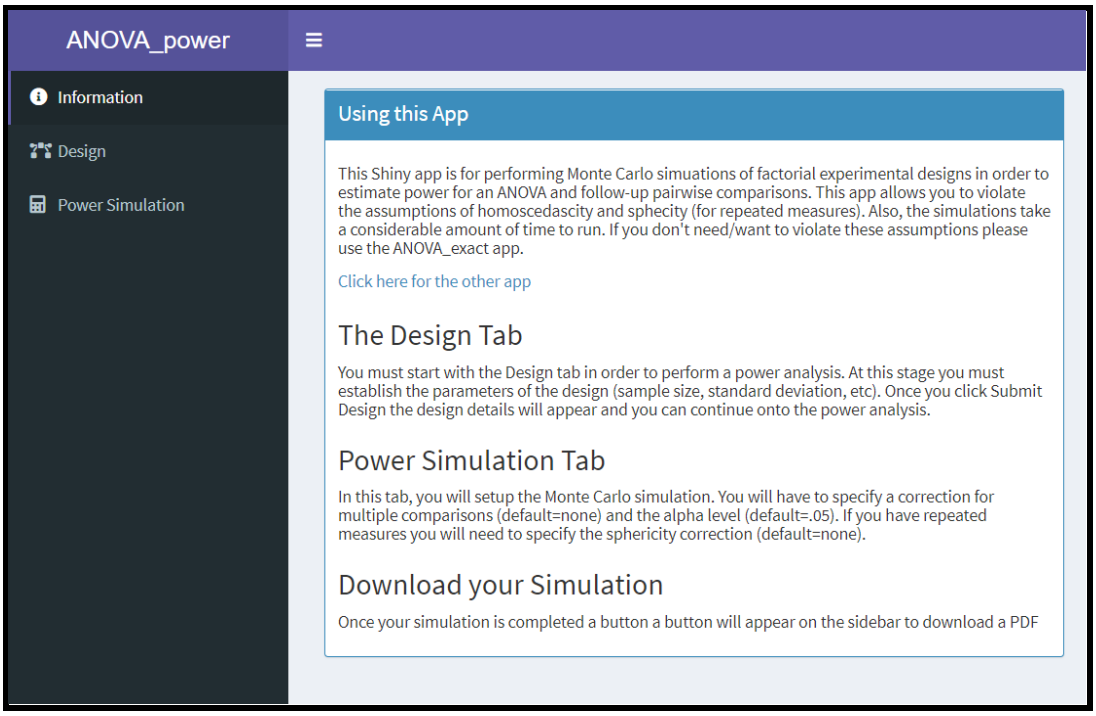
This design produced four unique groups to test the generative learning effect. There was a control group which neither produced nor received a drawing. Two groups received just one intervention, either producing a drawing or receiving an author-generated drawing. The final group received both interventions: producing *and* receiving a drawing.

The dependent variable was their score on a drawing test. The children completed the test in classrooms and independently read an instructional booklet. The children read the booklet and were supported by one of the four interventions described above. Finally, they completed a test of their comprehension about the topic through the medium of drawing the main concepts. The authors calculated the drawing score by adding up the correct components based on a coding scheme and converting them to z-scores. This allowed the authors to compare results across their studies as the tests used different numbers of questions.

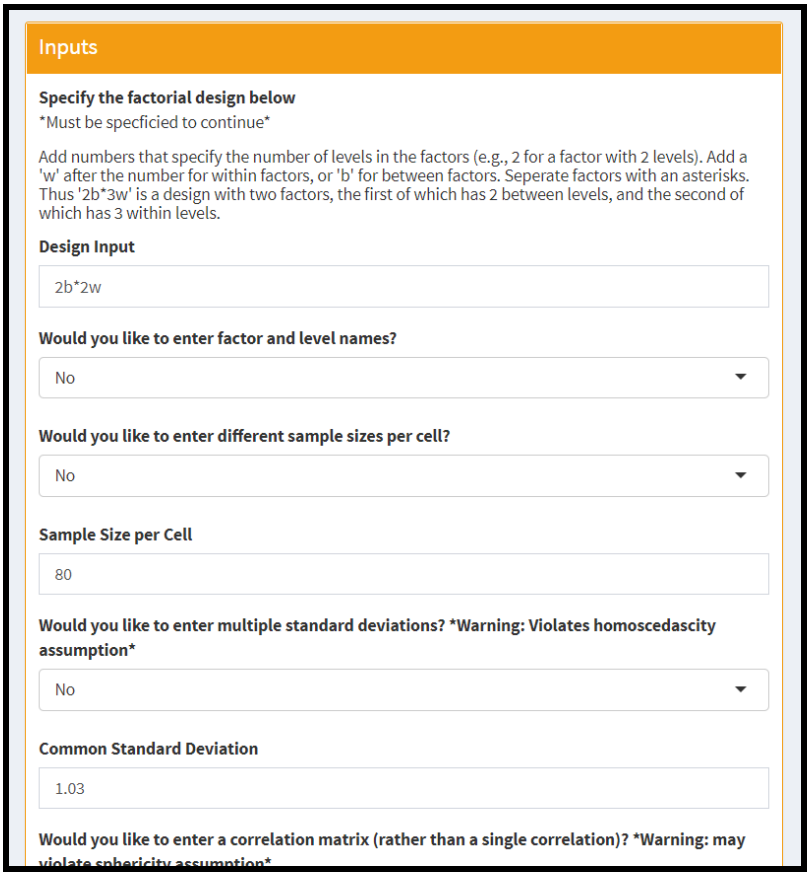
The researchers expected the combined learner- and author-generated group to score the highest and the control group who made no drawings to score the lowest. Both main effects and the interaction were significant in their study. We are interested in the interaction as we want to know whether comprehension changes across the four groups, not just isolated to author- or learner-generated effects.

We will reproduce the means from Schmeck et al. (2014) to see how many participants we would need to detect their effects. The combined (M = 0.66, SD = .22) and learner-generated groups (M = 0.63, SD = 0.19) scored the highest. The author-generated (M = 0.50, SD = 0.25) and control (M = 0.30, SD = 0.16) groups scored the lowest. This means we want to use the Superpower app to see how many participants we would need to detect these effects with our desired level of power.

Opening the Superpower app should present the following window:



To begin, we need the second tab to outline our design:



This provides us with several options to specify the design:

- **Design input:** This is where you specify your IVs and levels. For our example, we need to specify 2b*2b for a 2x2 between-subjects design. If you had a 3x2x2 between-subjects design, you would enter 3b*2b*2b.
- **Would you like to enter factor and level names?** Selecting yes will open a new input box for you to specify the names of your IVs and levels. This is highly recommended as it will make it easier to understand the output. For our example, enter: Author, Yes, No, Learner, Yes, No. The order is defined by the first IV name and each level, then the second IV name and each level, and so on.
- **Would you like to enter different sample sizes per cell?** Keeping the default no means you are defining the same sample size per group for equal group sizes. If you plan on having different group sizes, you can select yes and enter different group sizes.
- **Sample size per cell:** This is your sample size to enter per group. We will start with 20 in each group as this is historically a rule of thumb people have followed.
- **Would you like to enter multiple standard deviations?** If you expect different standard deviations for your groups, you can select yes. For our example, enter Yes as we have different standard deviations to reproduce their results. This will create a table to enter values rather than just one box. As we defined a 2x2 design, there are four cells. If we defined a 3x3 design, there would be nine cells. Pay attention to the order you enter the standard deviations as it should match the factors and levels you entered previously. a1_b1 is the first level of IV1 and the first level of IV2. a1_b2 is the first level of IV1 and the second level of IV2, and so on. Enter 0.22, 0.25, 0.19, and 0.16.
- **Means for Each Cell in the Design:** Finally, we can enter the means for each group and condition. Pay attention to the order you enter the means as it should match the factors and levels you entered previously. a1_b1 is the first level of IV1 and the first level of IV2. a1_b2 is the first level of IV1 and the second level of IV2, and so on. Enter 0.66, 0.50, 0.63, and 0.30.

Top tip: Specific to a factorial between-subjects design, if you do not know what means you expect per cell, you can use Cohen's d instead. If you set the standard deviation to 1, the difference in means across the cells are essentially expressed as standardised mean differences. For example, if you entered 0, 0.5, 0, 0.75, this would be equivalent to expecting a difference of Cohen's d = 0.5 between the first two groups, d = 0.75 between the second two groups, and d = 0.25 between the second and fourth group. This can be helpful if you know the standardised mean difference you want to power for, but not the means and standard deviations per cell.

Clicking Set up design will create a plot on the right side of the screen. This visualises the values you entered and it is useful for double checking you entered the values in the right order. If you followed the instructions, you should have the following screen which has been split into two images:

Inputs

Specify the factorial design below

Must be specficied to continue

Add numbers that specify the number of levels in the factors (e.g., 2 for a factor with 2 levels). Add a 'w' after the number for within factors, or 'b' for between factors. Seperate factors with an asterisks. Thus '2b*3w' is a design with two factors, the first of which has 2 between levels, and the second of which has 3 within levels.

Design Input

2b*2b

Would you like to enter factor and level names?

Yes

Specify one word for each factor (e.g., AGE and SPEED) and the level of each factor (e.g., old and yound for a factor age with 2 levels).

Factor & level labels

Author, Yes, No, Learner, Yes, No

Would you like to enter different sample sizes per cell?

No

Sample Size per Cell

20

Would you like to enter multiple standard deviations? *Warning: Violates homoscedascity assumption*

Yes

Specify the list of standard deviations.

	a1_b1	a1_b2	a2_b1	a2_b2
sd	0.22	0.25	0.19	0.16

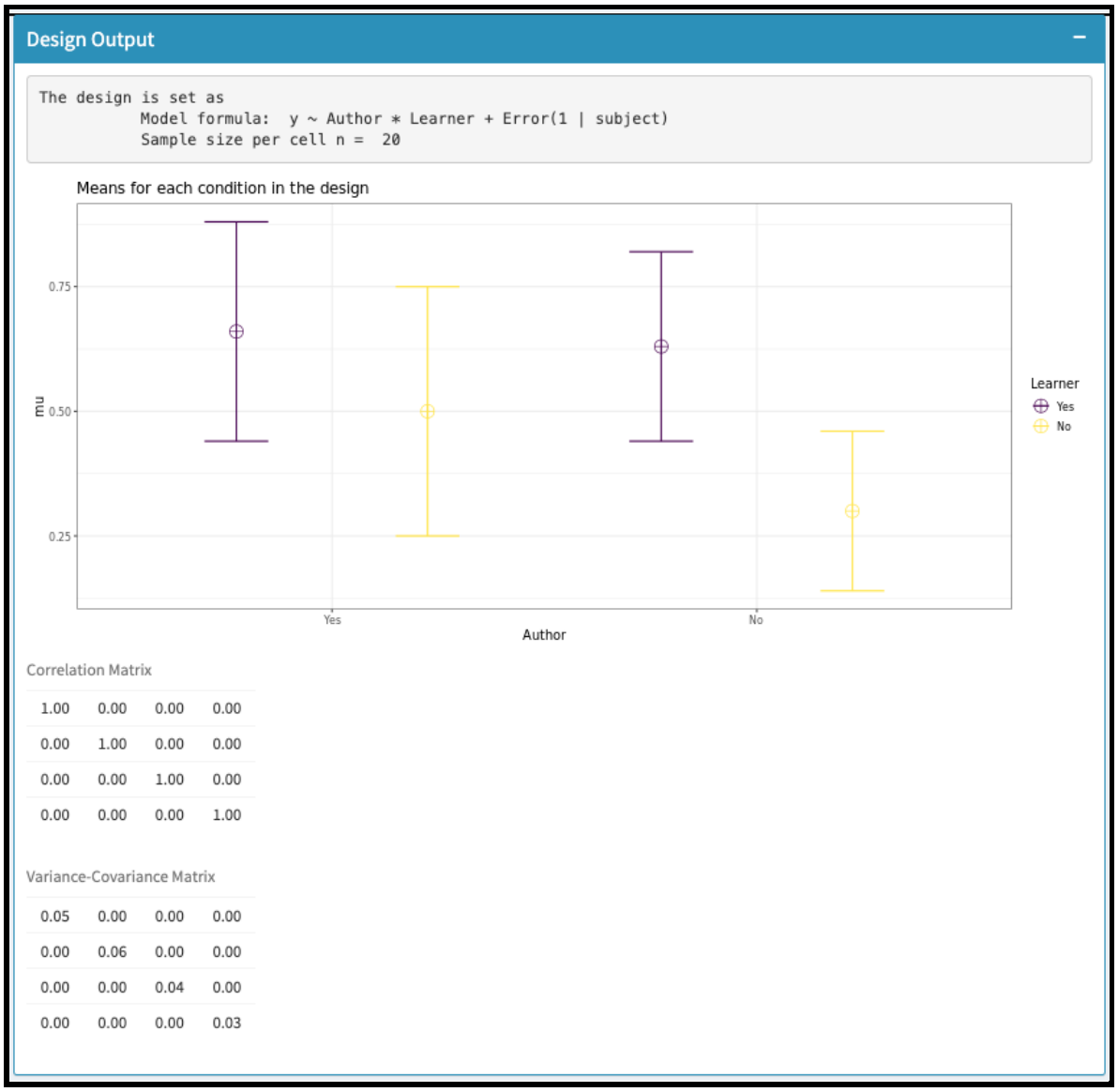
Note that for each cell in the design, a mean must be provided. Thus, for a '2b*3w' design, 6 means need to be entered. Means need to be entered in the correct order. The app provides a plot so you can check if you entered means correctly.

Means for Each Cell in the Design

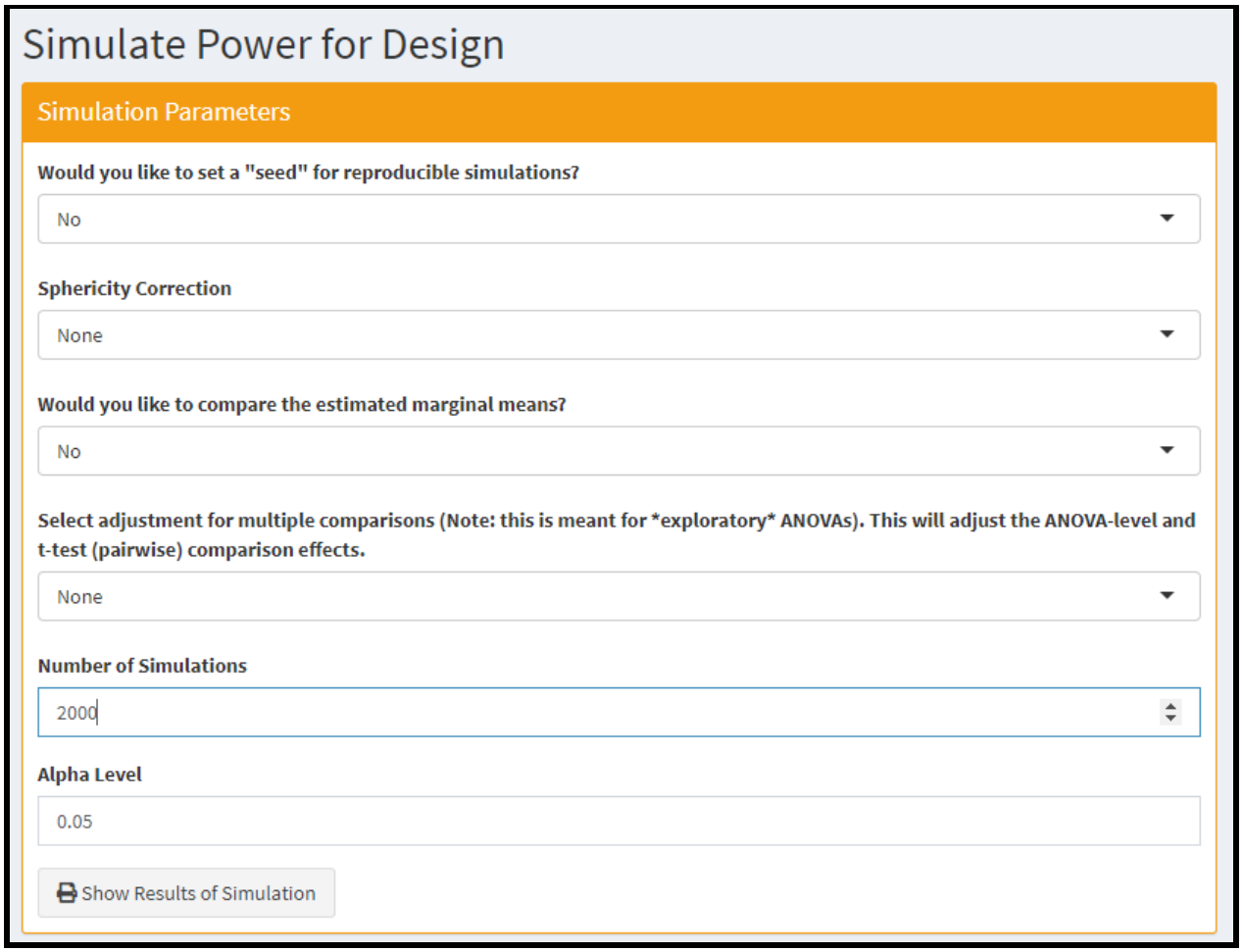
	a1_b1	a1_b2	a2_b1	a2_b2
mu	0.66	0.50	0.63	0.30

Click the button below to set up the design - Check the output to see if the design is as you intended, then you can run the simulation on the next tab.

Set-Up Design



Now you have set up the design, you need to click on the next tab: Power Simulation. If you populated your design in the previous tab, the power simulation tab will allow you to select the options to run the simulation. You should have the following options:

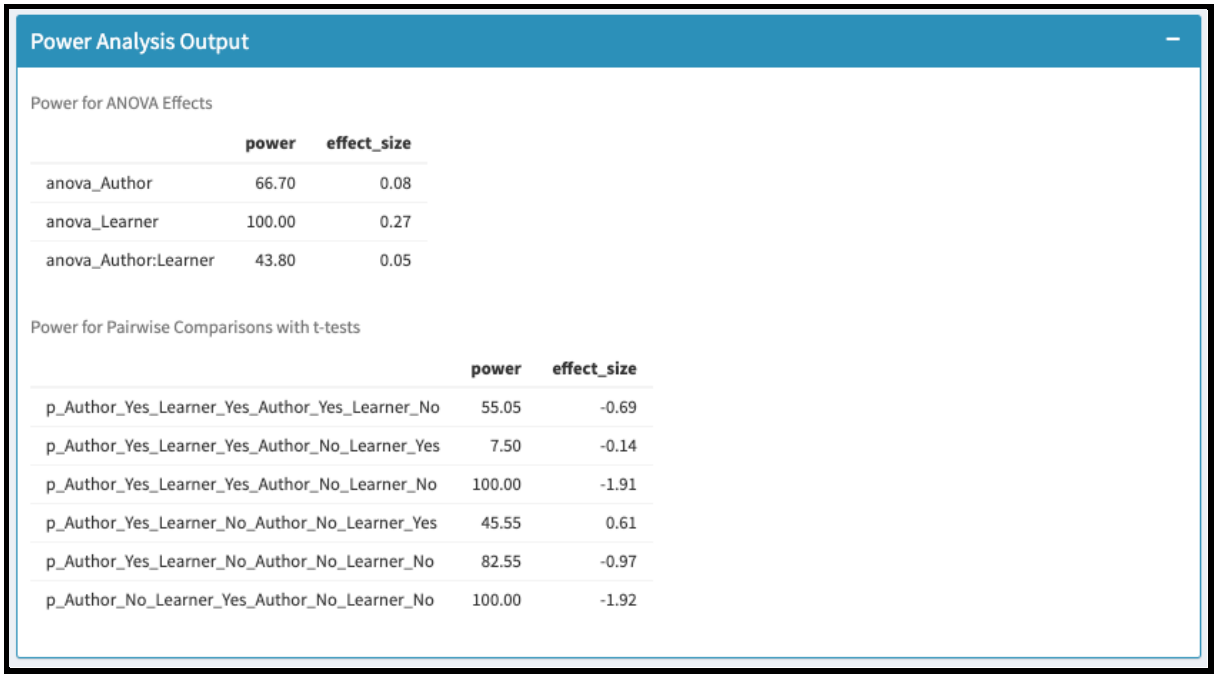


- **Would you like to set a "seed" for reproducible simulations?** If you select yes from the dropdown menu, it will create a new box to enter a seed. As simulations are based on semi-random numbers, setting a seed means the numbers will generate from the same starting point. This means you will get the same answer every time you run it with the same parameters. Not selecting a seed will mean the results will vary slightly each time. Select Yes and enter 2021.
- **Sphericity Correction**: Selecting Yes would correct the values for violating sphericity, but it does not apply for our example as we have a fully between-subjects design.
- **Would you like to compare the estimated marginal means?** Selecting Yes will create several new options to calculate the estimated marginal means and add corrections for multiple comparisons. We have a 2x2 design here, so keep it as No.
- **Select adjustment for multiple comparisons**: This provides you with the option to correct the ANOVA for multiplicity (increased type I error rate due to the large number of effects in an exploratory ANOVA). Keep it as None as we only have three effects in our 2x2 design.
- **Number of simulations**: This determines how many simulations your results are based on. When you are playing around with the settings, you can select a small number (e.g.,

200) so the simulations do not take very long. Once you are happy, you can select a larger number (e.g., 2000) so the results are more stable. Keep it as 2000 for now.

- **Alpha level**: We traditionally use an alpha value of .05 in psychology, but if you want to change it (e.g., to .005), you can change it here. For this example, we will keep it as .05.

Selecting these options and clicking show results of simulation should present you with the following results:



Power Analysis Output

Power for ANOVA Effects

	power	effect_size
anova_Author	66.70	0.08
anova_Learner	100.00	0.27
anova_Author:Learner	43.80	0.05

Power for Pairwise Comparisons with t-tests

	power	effect_size
p_Author_Yes_Learner_Yes_Author_Yes_Learner_No	55.05	-0.69
p_Author_Yes_Learner_Yes_Author_No_Learner_Yes	7.50	-0.14
p_Author_Yes_Learner_Yes_Author_No_Learner_No	100.00	-1.91
p_Author_Yes_Learner_No_Author_No_Learner_Yes	45.55	0.61
p_Author_Yes_Learner_No_Author_No_Learner_No	82.55	-0.97
p_Author_No_Learner_Yes_Author_No_Learner_No	100.00	-1.92

The power analysis is split between two tables. The first table is power for each effect in your ANOVA. In a 2x2 design, we have two main effects and one interaction. The second table is the pairwise comparisons of all the levels of your IVs. You have to decide which effects you are interested in to make sure your study is powered to detect them. As there are several effects in an ANOVA, you may not be interested in all of them. Therefore, you should power for the effects to address your hypotheses.

For this example, we are interested in the interaction as we expect the drawing scores to change across our four groups. Schmeck et al. (2014) predicted the combined group would score highest and the control would score lowest. They focused on all the pairwise comparisons:

- Combined vs author-generated
- Combined vs control
- Combined vs learner-generated
- Learner-generated vs author-generated
- Learner-generated vs control
- Author-generated vs control group

With 20 participants per group (N = 80), there was 66% power for the main effect of author, 100% for the main effect of learner, and 43% for the interaction. The pairwise comparisons ranged from 7.5% to 100% power. This means we would need more participants to power the effects of interest.

If you return to the design tab, you can explore alternative values for the sample size per group. This can take some trial and error as there is no power curve in the app. If you enter 50 participants per group (N = 200), this should provide you with 81% power for the interaction. In the pairwise comparisons, all the combinations reached at least 80% power apart from combined vs learner-generated. This is consistent with Schmeck et al. (2014), as the only non-significant pairwise comparison was between these two groups they expected to score the highest.

How can this be reported?
For this example, we could report it like this:

> "We conducted a simulated power analysis using the SuperPower Shiny app. We defined a between-subjects 2x2 design with one between-subjects IV of learner-generated drawing (yes vs no) and one between-subjects IV of author-generated drawing (yes vs no). Based on previous research, we assumed the combined (*M* = 0.66, *SD* = .22) and learner-generated groups (*M* = 0.63, *SD* = 0.19) would score the highest. The author-generated (*M* = 0.50, *SD* = 0.25) and control (*M* = 0.30, *SD* = 0.16) groups would score the lowest. We were interested in the interaction effect to evaluate whether the combined and learner-generated group would outperform the other groups, but not each other. Using 2000 simulations, we would need 50 participants per group (N = 200) to achieve 81% power for the interaction, meaning our target sample size is at least 50 participants per group."

In comparison to previous statements reporting a power analysis, this requires much more detail to be reproducible. These are the parameters you would need to enter into SuperPower to receive the same results.

Remember: in a factorial design you need to think about which effects you power your study for and you need to use a sample size that covers all your effects of interest at your desired level of power.

5.3. Factorial within-subjects design

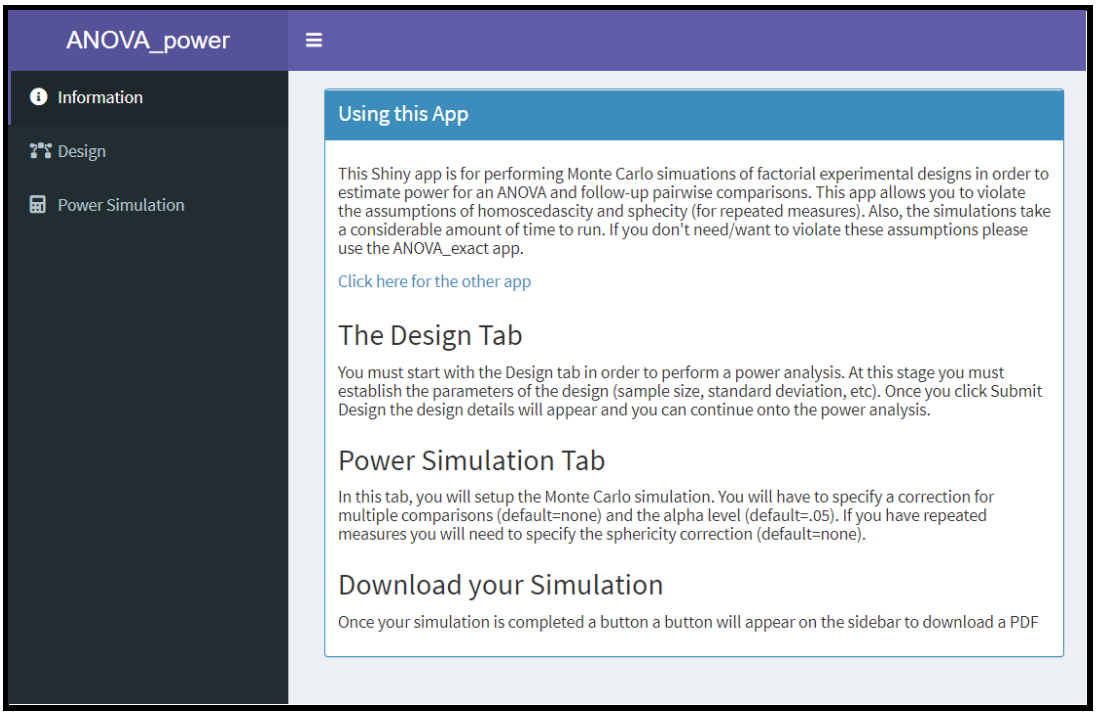
For this example, I will pretend we are going to replicate Rode and Ringel (2019). The authors were interested in how students' anxiety and confidence towards interpreting statistical output would change from the start to the end of a short statistics course. Participants completed a course either using R or SPSS and the researchers measured their anxiety and confidence towards both. They wanted to find out whether the course would decrease anxiety and increase confidence in both the software they learnt and the software they did not learn to see if the effects transferred. This means their original study was a 2x2x2 mixed design, but for the purposes of this example we will just focus on one piece of software to create a constrained 2x2 within-subjects design.

Our study will just teach students to use R and we are interested in how their anxiety towards statistics output changes from the start of the course to the end, for both R and SPSS output. This means we have one within-subjects IV of time: the start and the end of the course. We also have a within-subjects IV of software to interpret: R and SPSS. As in the original study, we will measure anxiety using the mean of two questions on a scale of 1-7. Lower scores indicate lower anxiety towards understanding statistics software.

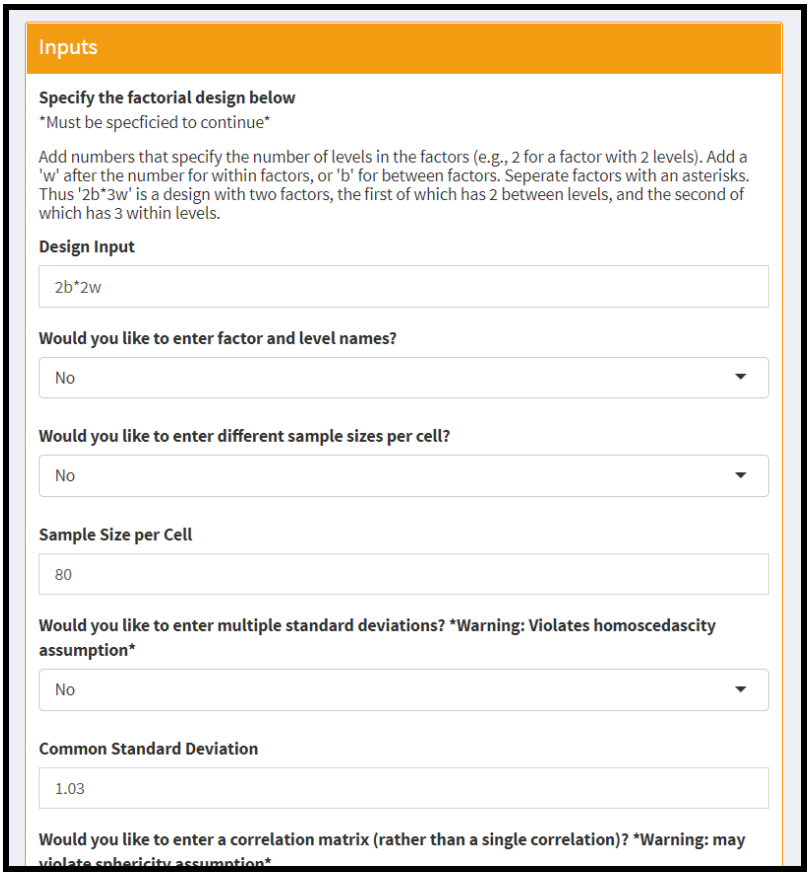
Based on Rode and Ringel (2019), we expect anxiety to decrease from the start of the course to the end of the course. This means we are interested in the main effect of time: does anxiety decrease from the start to the end of the course? We also expect this decrease to affect both pieces of software, but the decrease should be larger for the R condition since that is the software they learnt on the course. This means we are interested in the interaction between time and software to interpret: does anxiety decrease more in the R condition than in the SPSS condition?

We will reproduce the means from Rode and Ringel (2019) to see how many participants we would need to detect their effects. Anxiety decreased from an average of 5.00 to 2.00 for the R output. Anxiety decreased from 4.50 to 2.20 for the SPSS output. The authors did not report an SD for each condition, so we will use 1.5 as a convervative estimate from the values they did report. Finally, this study is one of the best examples for sharing the correlation matrix for repeated measures factors which we will need here. We will start off using a lenient estimate of r = .70; you could then explore how the sample size would change for more conservative estimates.

Opening the Superpower app should present the following window:



To begin, we need the second tab to outline our design:



This provides us with several options to specify the design:

- **Design input:** This is where you specify your IVs and levels. For our example, we need to specify 2w*2w for a 2x2 within-subjects design. If you had a 3x2x2 within-subjects design, you would enter 3w*2w*2w etc.
- **Would you like to enter factor and level names?** Selecting yes will open a new input box for you to specify the names of your IVs and levels. This is highly recommended as it will make it easier to understand the output. For our example, enter: Output, R, SPSS, Time, Start, End. The order is defined by the first IV name and each level, then the second IV name and each level, and so on.
- **Would you like to enter different sample sizes per cell?** Keeping the default no means you are defining the same sample size per group for equal group sizes. If you plan on having different group sizes, you can select yes and enter different group sizes. For this example, it would not apply as we only have within-subject IVs.
- **Sample size per cell:** This is your sample size to enter per group. We will start with 20 participants as this is historically a rule of thumb people have followed.
- **Would you like to enter multiple standard deviations?** If you expect different standard deviations for your groups, you can select yes. For our example, we are going to use one conservative estimate for the standard deviation of 1.5.
- **Would you like to enter a correlation matrix?** As we have within-subjects IVs, we need to take the correlation between conditions into account. This option would allow us to specify a whole correlation matrix for the combination of levels. This is potentially the most difficult parameter to choose as its rare for experimental studies to report the correlation between variables. Fortunately, Rode and Ringel (2019) did report the correlation matrix. For this example, we will choose an lenient value of r = .70. We can explore how power changes for more conservative estimates later.
- **Means for Each Cell in the Design:** Finally, we can enter the means for each group and condition. As we defined a 2x2 design, there are four cells. If we defined a 3x3 design, there would be nine cells. Pay attention to the order you enter the means as it should match the factors and levels you entered previously. a1_b1 is the first level of IV1 and the first level of IV2. a1_b2 is the first level of IV1 and the second level of IV2, and so on. Enter 5, 2, 4.5, 2.2. This means our two measurements for R are first and the two for SPSS are second.

Clicking Set up design will create a plot on the right. This visualises the values you entered and it is useful for double checking you entered the values in the right order. If you followed the instructions, you should have the following screen which can been split into two images:

Design Input

2w*2w

Would you like to enter factor and level names?

Yes

Specify one word for each factor (e.g., AGE and SPEED) and the level of each factor (e.g., old and yound for a factor age with 2 levels).

Factor & level labels

Output, R, SPSS, Time, Start, End

Would you like to enter different sample sizes per cell?

No

Sample Size per Cell

20

Would you like to enter multiple standard deviations? *Warning: Violates homoscedascity assumption*

No

Common Standard Deviation

1.5

Would you like to enter a correlation matrix (rather than a single correlation)? *Warning: may violate sphericity assumption*

No

Common correlation among within-subjects factors

0.5

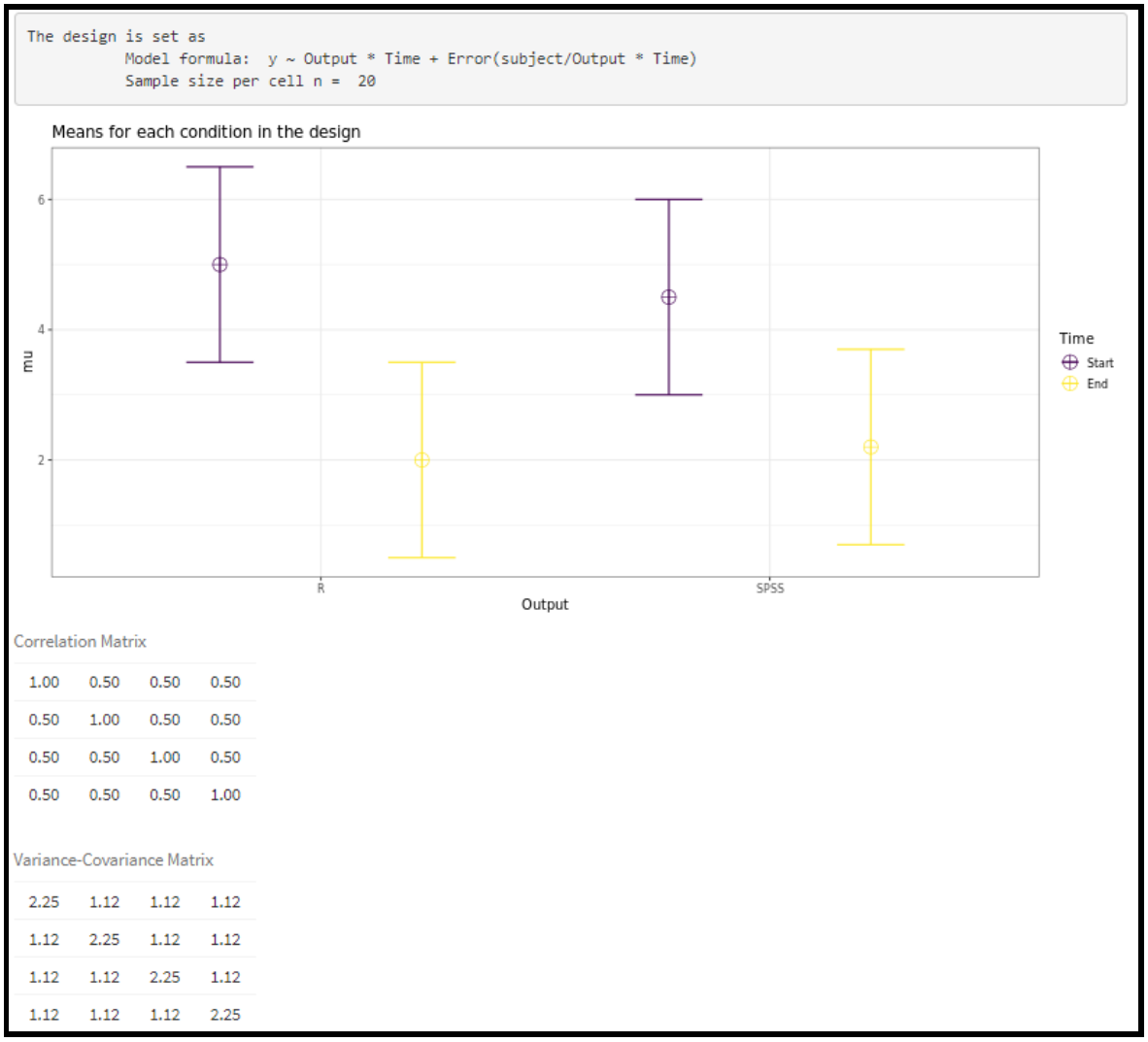
Note that for each cell in the design, a mean must be provided. Thus, for a '2b*3w' design, 6 means need to be entered. Means need to be entered in the correct order. The app provides a plot so you can check if you entered means correctly.

Means for Each Cell in the Design

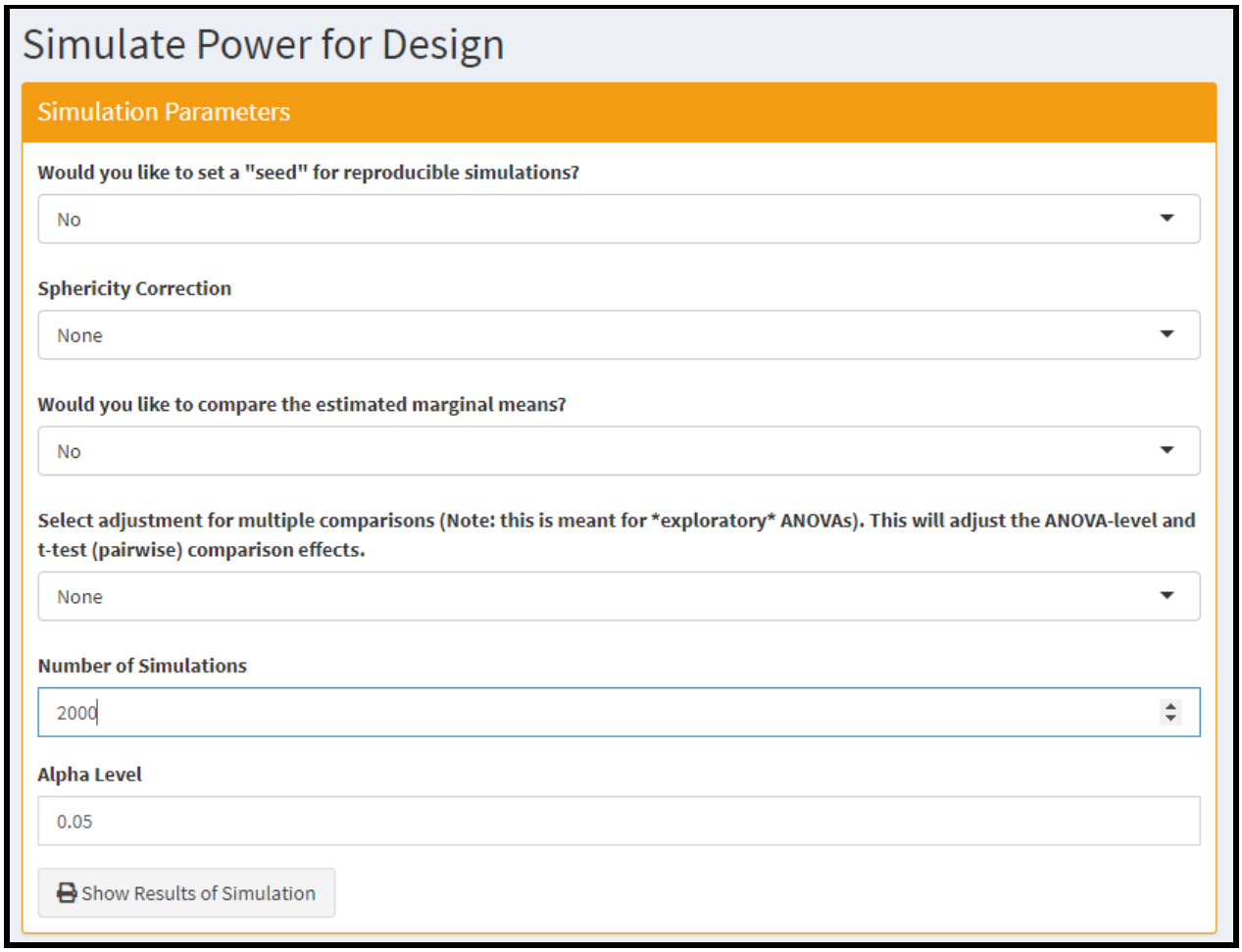
	a1_b1	a1_b2	a2_b1	a2_b2
mu	5	2	4.5	2.2

Click the button below to set up the design - Check the output to see if the design is as you intended, then you can run the simulation on the next tab.

Set-Up Design



Now you have set up the design, you need to click on the next tab: Power Simulation. If you populated your design in the previous tab, the power simulation tab will allow you to select the options to run the simulation. You should have the following options:



- **Would you like to set a "seed" for reproducible simulations?** If you select yes from the dropdown menu, it will create a new box to enter a seed. As simulations are based on semi-random numbers, setting a seed means the numbers will generate from the same starting point. This means you will get the same answer every time you run it with the same parameters. Not selecting a seed will mean the results will vary slightly each time. Select Yes and enter 2021.
- **Sphericity Correction**: Selecting Yes would correct the values for violating sphericity, but it does not apply for our example as we did not set the correlations to vary by condition.
- **Would you like to compare the estimated marginal means?** Selecting Yes will create several new options to calculate the estimated marginal means and add corrections for multiple comparisons. We have a 2x2 design here, so keep it as No.
- **Select adjustment for multiple comparisons**: This provides you with the option to correct the ANOVA for multiplicity (increased type I error rate due to the large number of effects in an exploratory ANOVA). Keep it as None as we only have three effects in our 2x2 design.

- **Number of simulations**: This determines how many simulations your results are based on. When you are playing around with the settings, you can select a small number (e.g., 200) so the simulations do not take very long. Once you are happy, you can select a larger number (e.g., 2000) so the results are more stable. Keep it as 2000 for now.
- **Alpha level**: We traditionally use an alpha value of .05 in psychology, but if you want to change it (e.g., to .005), you can change it here. For this example, we will keep it as .05.

Selecting these options and clicking show results of simulation should present you with the following results:

Power for ANOVA Effects

	power	effect_size
anova_Output	9.70	0.07
anova_Time	100.00	0.87
anova_Output:Time	29.30	0.14

Power for Pairwise Comparisons with t-tests

	power	effect_size
p_Output_R_Time_Start_Output_R_Time_End	100.00	-2.09
p_Output_R_Time_Start_Output_SPSS_Time_Start	28.35	-0.34
p_Output_R_Time_Start_Output_SPSS_Time_End	100.00	-1.94
p_Output_R_Time_End_Output_SPSS_Time_Start	100.00	1.75
p_Output_R_Time_End_Output_SPSS_Time_End	9.70	0.14
p_Output_SPSS_Time_Start_Output_SPSS_Time_End	100.00	-1.60

The power analysis is split between two tables. The first table is power for each effect in your ANOVA. In a 2x2 design, we have two main effects and one interaction. The second table is the pairwise comparisons of all the levels of your IVs. You have to decide which effects you are interested in to make sure your study is powered to detect them. As there are several effects in an ANOVA, you may not be interested in all of them like the main effect of output in this example. Therefore, you should power for the effects you are interested in to address your hypotheses.

For this example, we are interested in the main effect of time and potentially the interaction. We expect anxiety to decrease for both output conditions, but we expect a larger decrease for the R condition as that is what the students learnt on the course. With 20 participants, we already have 100% power for the main effect of time and 29% for the interaction. Both of the pairwise comparisons comparing start vs end for each type of output has 100% power, but its the

interaction which would tell us whether this decrease is moderated by the type of output. This means we will need more participants.

If you return to the design tab, you can explore alternative values for the sample size. This can take some trial and error as there is no power curve in the app. If you enter 71 participants, this should provide you with approximately 80% power for the interaction effect between output and time. Since all the pairwise comparisons were covered by the original power analysis with 20 participants, you would need to decide whether the interaction is worth powering with the anticipated effects. You can also explore to see how decreasing the standard deviation and increasing the correlation affects power.

How can this be reported?
For this example, we could report it like this:

> "We conducted a simulated power analysis using the SuperPower Shiny app. We defined a 2x2 within-subjects design with one IV of software output (R and SPSS) and one IV of time (start and end of the course). We assumed anxiety towards R would decrease from 5 to 2, and it would decrease from 4.5 to 2.2 for SPSS output. We assumed a standard deviation of 1.5 around these means and a correlation of r = .70 between conditions. We were interested in the main effect of time and potentially the interaction. Using 2000 simulations, we would need 20 participants to achieve at least 80% power for the main effect of time and 71 participants to achieve 80% power for the interaction, meaning our target sample size is at least 71 participants."

In comparison to previous statements reporting a power analysis, this requires much more detail to be reproducible. These are the parameters you would need to enter into SuperPower to receive the same results.

Remember: in a factorial design you need to think about which effects you power your study for and you need to use a sample size that covers all your effects of interest at your desired level of power.

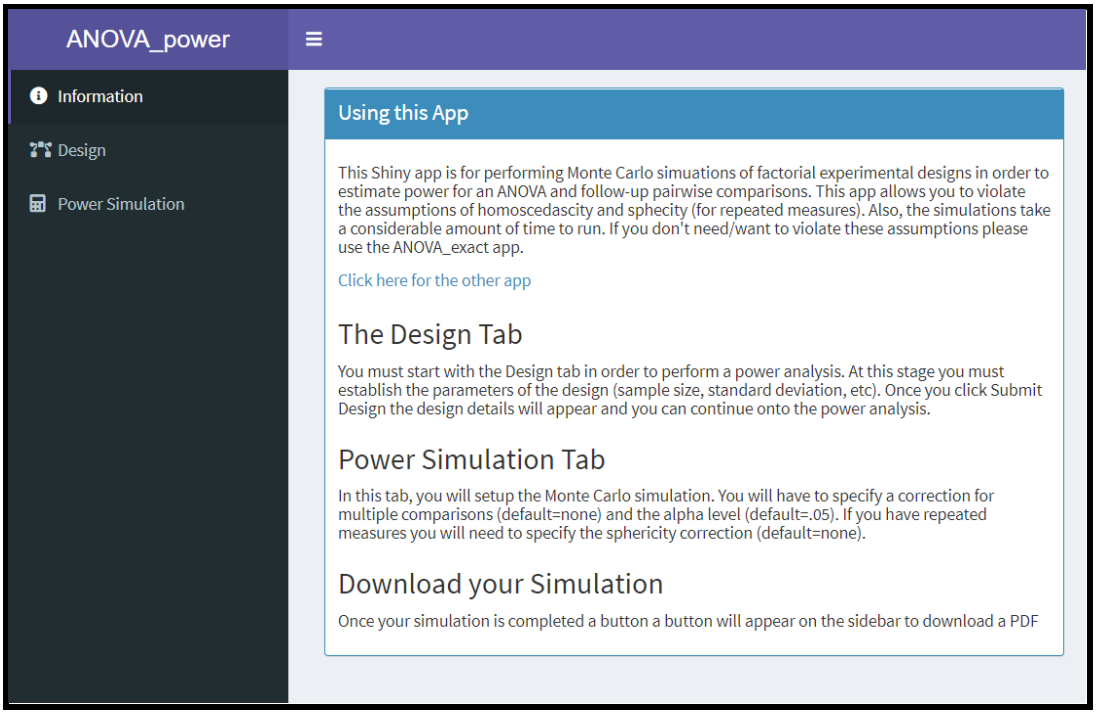
5.4. Factorial mixed design

For this example, I will use a study from my PhD (Bartlett et al., 2020). This study used a 2x2 design with one between-subjects IV of smoking group (daily smokers and non-daily smokers) and one within-subjects IV of presentation time (200ms and 500ms). We investigated something called attentional bias towards smoking cues and used a measure called the visual probe task. The idea is participants can respond faster to probes that replace smoking images compared to neutral images. This means positive values indicate attentional bias to smoking images and negative values indicate attentional bias to neutral images.

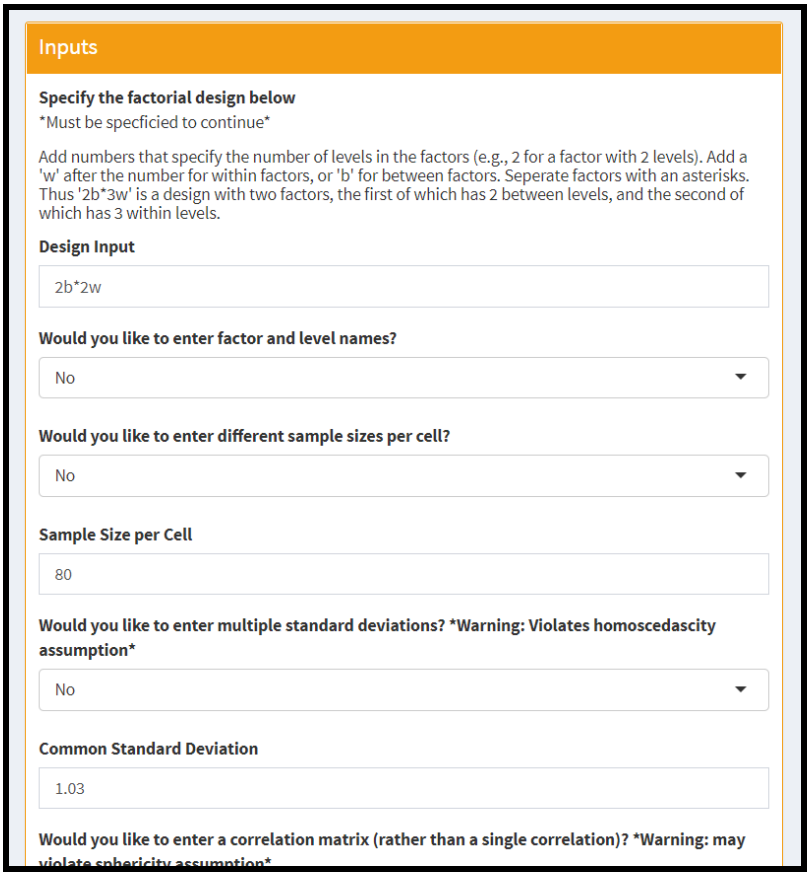
Based on previous research, we expected non-daily smokers to display greater attentional bias than daily smokers. We also expected a slightly larger effect in the 500ms condition than the 200ms condition. Therefore, our smallest effect sizes of interest were a 5ms difference in the 200ms condition and a 10ms difference in the 500ms condition. We used a conservative estimate for the standard deviation around these differences of 20ms.

We were interested in the main effect of smoking group (non-daily smokers to show a larger positive effect across both presentation time conditions) or an interaction (non-daily smokers to show greater attentional bias to smoking cues, but a larger effect in the 500ms condition). This means we want to use the Superpower app to see how many participants we would need to detect these effects with our desired level of power.

Opening the Superpower app should present the following window:



To begin, we need the second tab to outline our design:



This provides us with several options to specify the design:

- **Design input:** This is where you specify your IVs and levels. For our example, we need to specify 2b*2w for a 2x2 mixed design. If you had a 3x2x2 mixed design, you would enter 3b*2w*2w etc.
- **Would you like to enter factor and level names?** Selecting yes will open a new input box for you to specify the names of your IVs and levels. This is highly recommended as it will make it easier to understand the output. For our example, enter: Smoking_group, Daily, Nondaily, Presentation_time, 200ms, 500ms. The order is defined by the first IV name and each level, then the second IV name and each level, and so on.
- **Would you like to enter different sample sizes per cell?** Keeping the default no means you are defining the same sample size per group for equal group sizes. If you plan on having different group sizes, you can select yes and enter different group sizes.
- **Sample size per cell:** This is your sample size to enter per group. We will start with 20 in each group as this is historically a rule of thumb people have followed.
- **Would you like to enter multiple standard deviations?** If you expect different standard deviations for your groups, you can select yes. For our example, we are going to use one conservative estimate for the standard deviation of 20.

- **Would you like to enter a correlation matrix?** As we have a within-subjects IV, we need to take the correlation between conditions into account. This option would allow us to specify a whole correlation matrix for the combination of levels. This is potentially the most difficulty parameter to choose as its rare for experimental studies to report the correlation between variables. For this example, we will choose an optimistic value of r = .70. We can explore how power changes for more conservative estimates later.
- **Means for Each Cell in the Design:** Finally, we can enter the means for each group and condition. As we defined a 2x2 design, there are four cells. If we defined a 3x3 design, there would be nine cells. Pay attention to the order you enter the means as it should match the factors and levels you entered previously. a1_b1 is the first level of IV1 and the first level of IV2. a1_b2 is the first level of IV1 and the second level of IV2, and so on. Enter 0, 0, 5, 10. This means we expect a 5ms effect for non-daily smokers in the 200ms condition and a 10ms effect in the 500ms condition.

Clicking Set up design will create a plot on the right. This visualises the values you entered and it is useful for double checking you entered the values in the right order. If you followed the instructions, you should have the following screen which can been split into two images:

Inputs

Specify the factorial design below

Must be specficied to continue

Add numbers that specify the number of levels in the factors (e.g., 2 for a factor with 2 levels). Add a 'w' after the number for within factors, or 'b' for between factors. Seperate factors with an asterisks. Thus '2b*3w' is a design with two factors, the first of which has 2 between levels, and the second of which has 3 within levels.

Design Input

2b*2w

Would you like to enter factor and level names?

Yes

Specify one word for each factor (e.g., AGE and SPEED) and the level of each factor (e.g., old and yound for a factor age with 2 levels).

Factor & level labels

Smoking_group, Daily, Nondaily, Presentation_time, 200ms, 500ms

Would you like to enter different sample sizes per cell?

No

Sample Size per Cell

20

Would you like to enter multiple standard deviations? *Warning: Violates homoscedascity assumption*

No

Common Standard Deviation

20

Would you like to enter a correlation matrix (rather than a single correlation)? *Warning: may violate sphericity assumption*

No

Common correlation among within-subjects factors

0.7

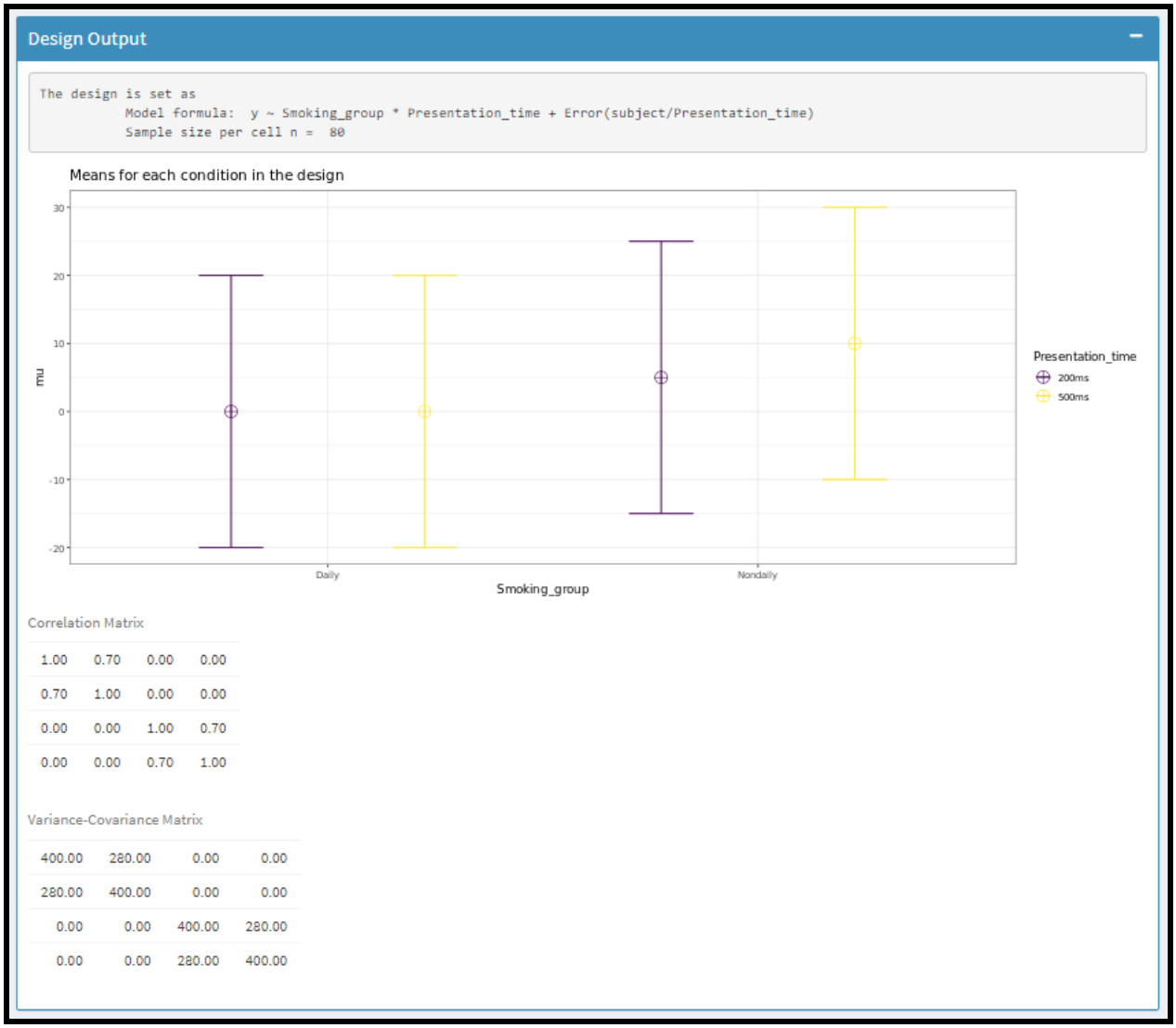
Note that for each cell in the design, a mean must be provided. Thus, for a '2b*3w' design, 6 means need to be entered. Means need to be entered in the correct order. The app provides a plot so you can check if you entered means correctly.

Means for Each Cell in the Design

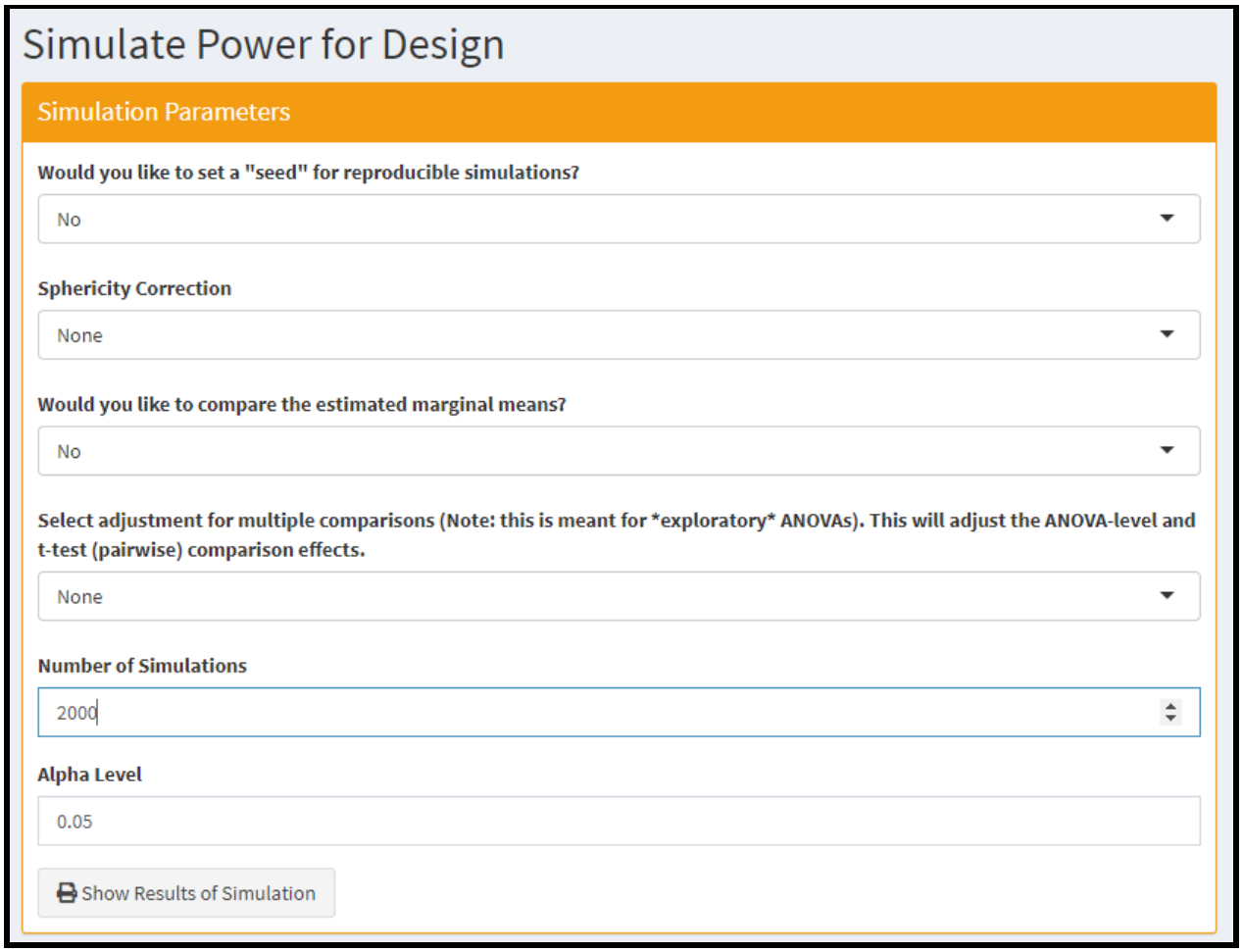
	a1_b1	a1_b2	a2_b1	a2_b2
mu	0	0	5	10

Click the button below to set up the design - Check the output to see if the design is as you intended, then you can run the simulation on the next tab.

Set-Up Design



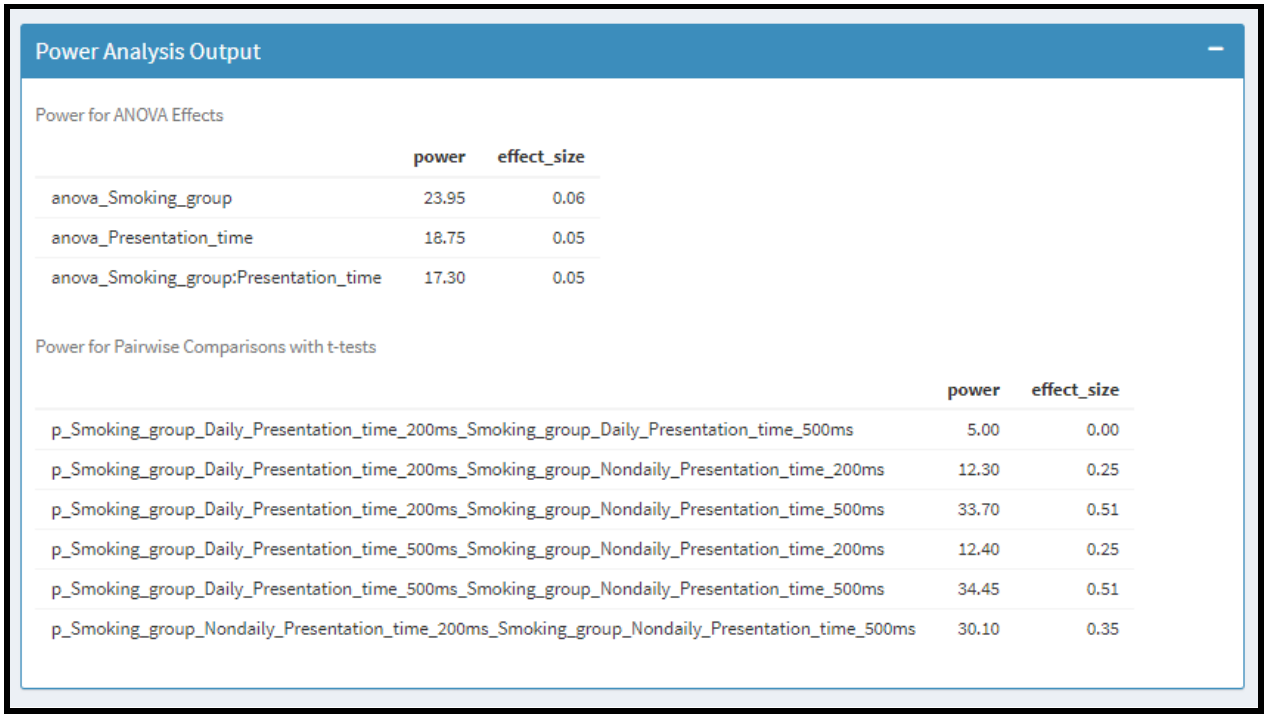
Now you have set up the design, you need to click on the next tab: Power Simulation. If you populated your design in the previous tab, the power simulation tab will allow you to select the options to run the simulation. You should have the following options:



- **Would you like to set a "seed" for reproducible simulations?** If you select yes from the dropdown menu, it will create a new box to enter a seed. As simulations are based on semi-random numbers, setting a seed means the numbers will generate from the same starting point. This means you will get the same answer every time you run it with the same parameters. Not selecting a seed will mean the results will vary slightly each time. Select Yes and enter 2021.
- **Sphericity Correction**: Selecting Yes would correct the values for violating sphericity, but it does not apply for our example as we did not set the correlations to vary by condition.
- **Would you like to compare the estimated marginal means?** Selecting Yes will create several new options to calculate the estimated marginal means and add corrections for multiple comparisons. We have a 2x2 design here, so keep it as No.
- **Select adjustment for multiple comparisons**: This provides you with the option to correct the ANOVA for multiplicity (increased type I error rate due to the large number of effects in an exploratory ANOVA). Keep it as None as we only have three effects in our 2x2 design.

- **Number of simulations**: This determines how many simulations your results are based on. When you are playing around with the settings, you can select a small number (e.g., 200) so the simulations do not take very long. Once you are happy, you can select a larger number (e.g., 2000) so the results are more stable. Keep it as 2000 for now.
- **Alpha level**: We traditionally use an alpha value of .05 in psychology, but if you want to change it (e.g., to .005), you can change it here. For this example, we will keep it as .05.

Selecting these options and clicking show results of simulation should present you with the following results:



Power Analysis Output

Power for ANOVA Effects

	power	effect_size
anova_Smoking_group	23.95	0.06
anova_Presentation_time	18.75	0.05
anova_Smoking_group:Presentation_time	17.30	0.05

Power for Pairwise Comparisons with t-tests

	power	effect_size
p_Smoking_group_Daily_Presentation_time_200ms_Smoking_group_Daily_Presentation_time_500ms	5.00	0.00
p_Smoking_group_Daily_Presentation_time_200ms_Smoking_group_Nondaily_Presentation_time_200ms	12.30	0.25
p_Smoking_group_Daily_Presentation_time_200ms_Smoking_group_Nondaily_Presentation_time_500ms	33.70	0.51
p_Smoking_group_Daily_Presentation_time_500ms_Smoking_group_Nondaily_Presentation_time_200ms	12.40	0.25
p_Smoking_group_Daily_Presentation_time_500ms_Smoking_group_Nondaily_Presentation_time_500ms	34.45	0.51
p_Smoking_group_Nondaily_Presentation_time_200ms_Smoking_group_Nondaily_Presentation_time_500ms	30.10	0.35

The power analysis is split between two tables. The first table is power for each effect in your ANOVA. In a 2x2 design, we have two main effects and one interaction. The second table is the pairwise comparisons of all the levels of your IVs. You have to decide which effects you are interested in to make sure your study is powered to detect them. As there are several effects in an ANOVA, you may not be interested in all of them like the main effect of presentation time in this example. Therefore, you should power for the effects you are interested in to address your hypotheses.

For this example, we are interested in the main effect of smoking group and potentially the interaction. We expect non-daily smokers to show greater attentional bias than daily smokers, and for this effect to apply to both presentation time conditions. This means in the pairwise comparisons table, we are interested in power for rows 2 (comparing each smoking group for the 200ms condition) and 5 (comparing each smoking group for the 500ms conditio). With 20

participants per group, we have 24% power for the main effect of smoking group and 17% for the interaction. This means we will need more participants.

If you return to the design tab, you can explore alternative values for the sample size per group. This can take some trial and error as there is no power curve in the app. If you enter 100 participants per group (N = 200), this should provide you with 82% power for the main effect of smoking group. Power for the interaction is still low at 60%. This means we have covered our first main effect of interest and you would need to decide whether the interaction is worth powering with the anticipated effects (if you keep on exploring, it would take around 155 participants per group: N = 310).

How can this be reported?
For this example, we could report it like this:

> "We conducted a simulated power analysis using the SuperPower Shiny app. We defined a mixed 2x2 design with one between-subjects IV of smoking group (daily smokers vs non-daily smokers) and one within-subjects IV of presentation time (200ms vs 500ms). We assumed non-daily participants would have a 5ms higher response time than daily smokers in the 200ms condition and a 10ms higher response time in the 500ms condition. We assumed a standard deviation of 20ms around these mean differences and a correlation of r = .70 between conditions. We were interested in the main effect of smoking group and potentially the interaction. Using 2000 simulations, we would need 100 participants per group (N = 200) to achieve 80% power for the main effect of smoking group and 155 per group (N = 310) to achieve 80% power for the interaction, meaning our target sample size is at least 155 participants per group."

In comparison to previous statements reporting a power analysis, this requires much more detail to be reproducible. These are the parameters you would need to enter into SuperPower to receive the same results.

Remember: in a factorial design you need to think about which effects you power your study for and you need to use a sample size that covers all your effects of interest at your desired level of power.

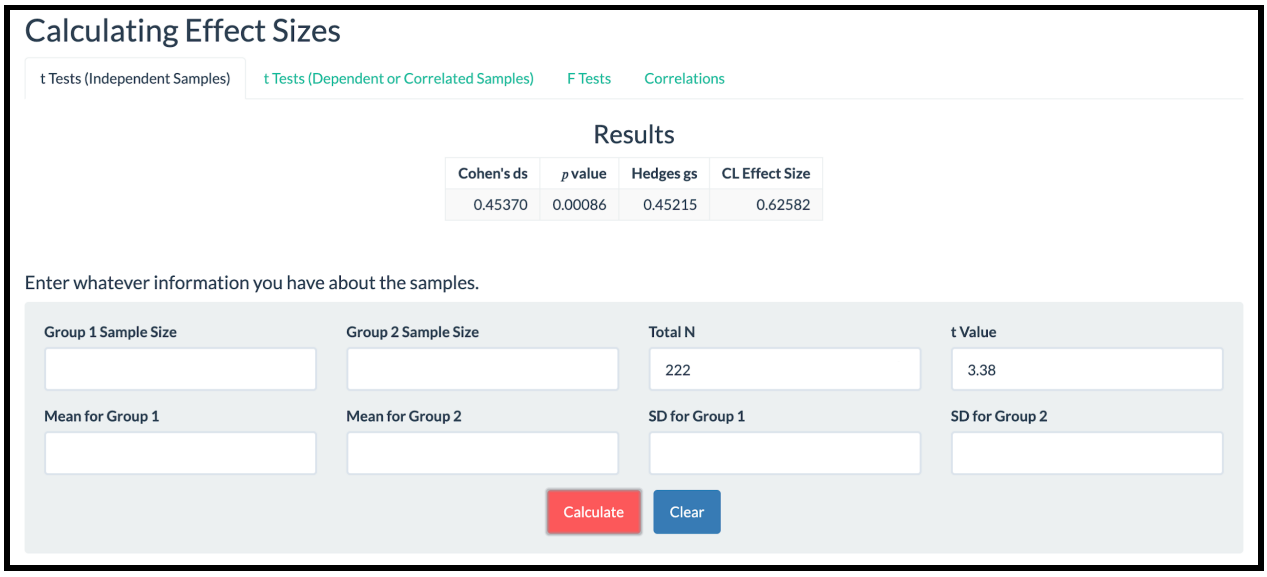
6. Sequential analysis

Alternatively, another way to efficiently work out the sample size is to check on your results as you are collecting data. You might not have a fully informed idea of the effect size you are expecting, or you may want to stop the study half way through if you already have convincing evidence. However, this must be done extremely carefully. If you keep collecting data and testing to see whether your results are significant, this drastically increases the type I error rate (Simons, Nelson, and Simonsohn, 2011). If you check enough times, your study will eventually produce a significant *p* value by chance even if the null hypothesis was really true. In order to check your results before collecting more data, you need to perform a process called sequential analysis (Lakens, 2014). This means that you can check the results intermittently, but for each time you check the results you must perform a type I error correction. This works like a Bonferroni correction for pairwise comparisons. For one method of sequential analysis, if you check the data twice, your alpha would be .029 instead of .05 in order to control the increase in type I error rate. This means for both the first and second look at the data, you would use an alpha value of .029 instead of .05. See Lakens (2014) for an overview of this process.

7. How to calculate an effect size from a test statistic

Throughout this guide, we have used effect size guidelines or meta-analytic effect sizes in order to select an effect size. However, these may not be directly applicable to the area of research you are interested in. You may want to replicate or extend an article you have read. One of the problems you may encounter with this, especially in older articles, is effect sizes are not consistently reported. This is annoying, but fortunately you can recalculate effect sizes from other information available to you, such as the test statistic and sample size. There is a direct relationship between the test statistic and effect size. This means if you have access to the sample size and test statistic, then you can recalculate an effect size based on this. You can also use descriptive statistics, but these may not be reported for each analysis. Due to APA style reporting, the test statistic and sample size should always be reported. There is a handy online app created by Katherine Wood for calculating effect sizes from the information available to you. We will be using this for the examples below.

For the first example, we will recalculate the effect size from Diemand-Yauman et al. (2011) as they report a t-test with Cohen's d, so we can see how well the recalculated effect size fits in with what they have reported. On page three of their article, there is the following sentence: "An independent samples t-test revealed that this trend was statistically significant ($t(220) = 3.38$, $p < .001$, Cohen's $d = 0.45$)". From the methods, we know there are 222 participants, and the t statistic equals 3.38. We can use this information in the online app to calculate the effect size. Select the independent samples t-test tab, and enter 222 into total N and 3.38 into t Value. If you click calculate, you should get the following output:

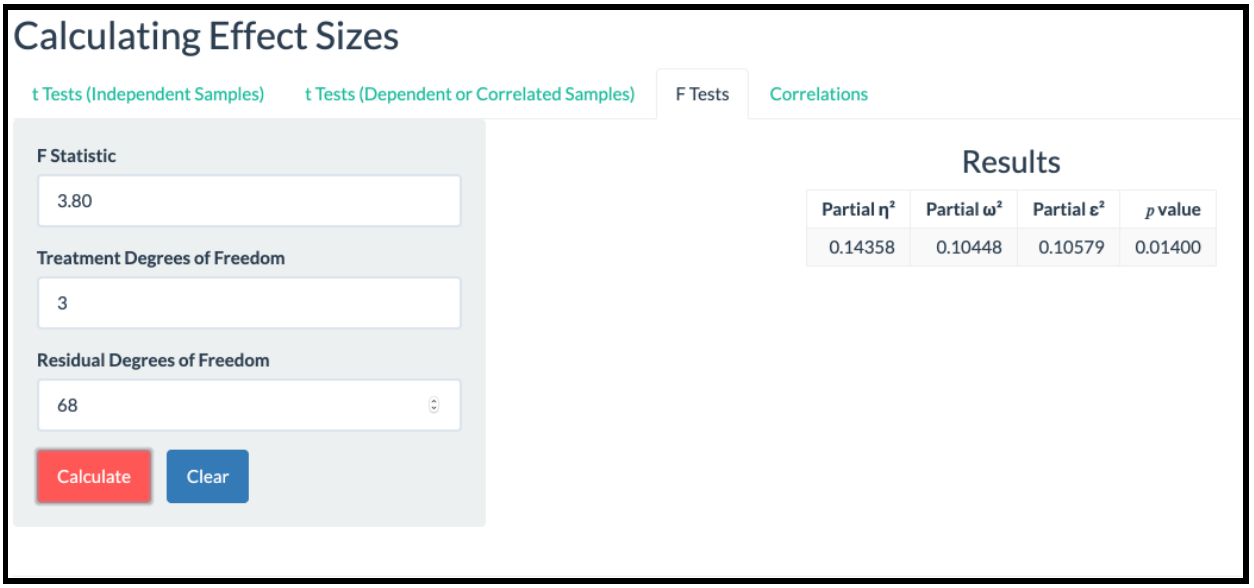


This shows that the recalculated effect size is nice and consistent. The value is 0.45 in both the article and our recalculation. This window shows the range of information that you can enter to

recalculate the effect size. The minimum information you need is the total N and t Value, or you will get an error message.

If you needed to recalculate the effect size for a paired samples t-test, then the options look very similar. You only have one sample size to think about, as each participant will complete both conditions. Therefore, we will move on to recalculating an effect size for ANOVA. Please note, this calculator only works for between-subjects ANOVA, including main effects and interactions. If you need the effect size for a one-way within-subjects ANOVA or a factorial mixed ANOVA, then you would need the full SPSS output, which is not likely to be included in an article. If it is available, you can use this spreadsheet by Daniël Lakens to calculate the effect size, but it's quite a lengthy process.

For the next example, we will use a one-way ANOVA reported in James et al. (2015). On page 1210, there is the following sentence: "there was a significant difference between groups in overall intrusion frequency in daily life, $F(3, 68) = 3.80$, $p = .01$, $\eta^2_p = .14$". If we click on the F tests tab of the online calculator, we can enter 3.80 for the F statistic, 3 for treatment degrees of freedom, and 68 for residual degrees of freedom. You should get the following output:



This shows that we get the same estimate for η^2_p as what is reported in the original article. Both values are .14.

8. How to increase statistical power

8.1. Increase the number of participants

As we have seen throughout this guide, one of the most straightforward ways of increasing statistical power is sampling more participants. If you have the resources available, performing a power analysis will hopefully allow you to target the sample size required for your smallest effect size of interest and your design. However, this is not always possible. If you are performing a student project and have little time or money, you might not be able to recruit a large sample. This means you will have to think of alternative strategies to perform an informative experiment.

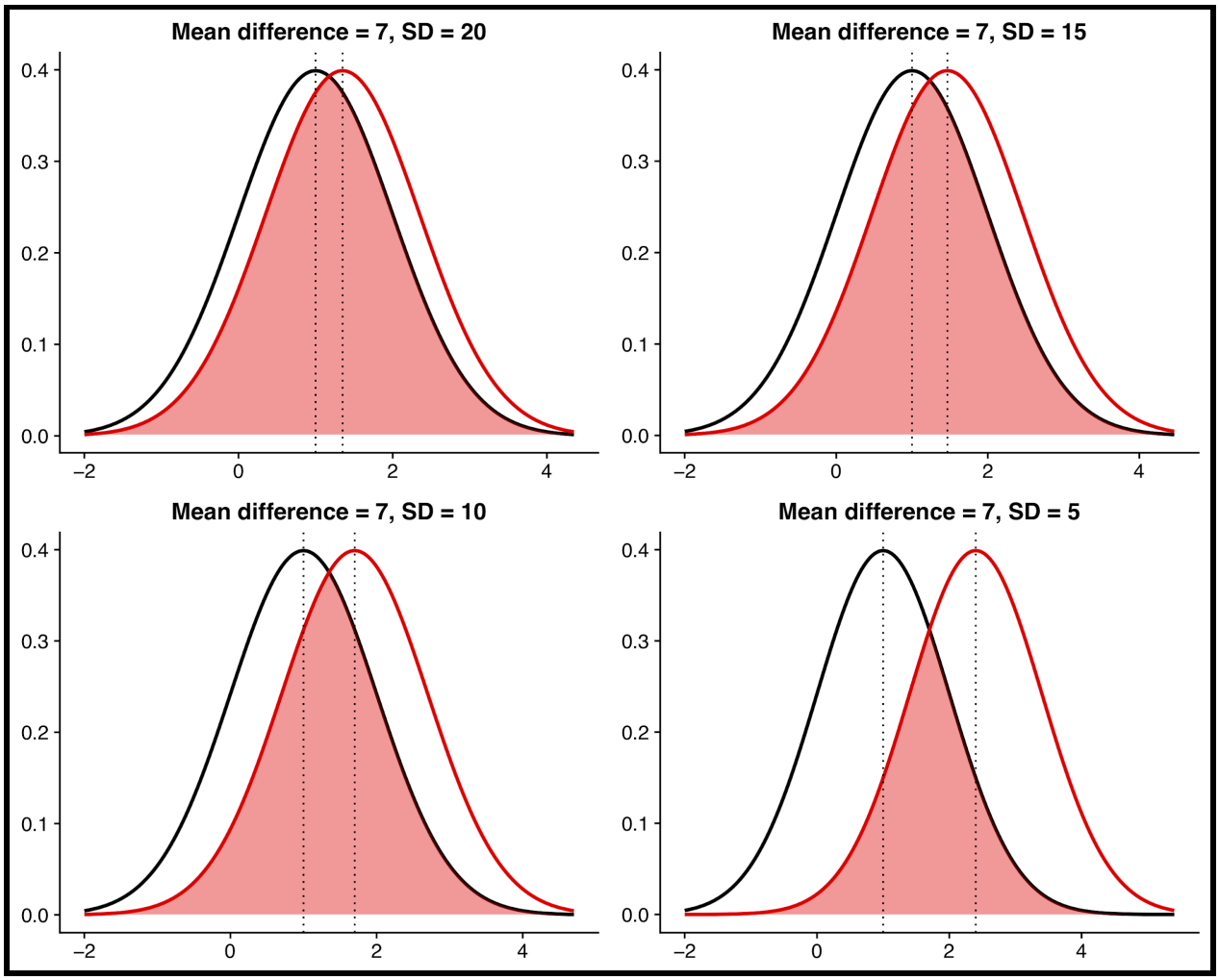
8.2. Increase the effect size

One alternative is to increase the unstandardised effect size. If you are manipulating your IV, you could increase the effect by increasing the dose or exposure. For example, if you were interested in the effect of alcohol on a certain behaviour, using a moderate dose of alcohol should have a larger effect than a small dose of alcohol in comparison to a no alcohol control group.

8.3. Decrease the variability of your effect

It may not always be possible to just increase the unstandardised effect size. Perhaps you are observing a behaviour rather than manipulating it. This is where you can increase the standardised effect size by making it easier to detect. In G*Power, we have been using Cohen's d, which is the standardised mean difference. This is the difference between groups or conditions divided by the standard deviation. This means that the difference is converted to a uniform scale expressed in standard deviations. For example, say we wanted to compare two groups on a reaction time experiment. There is a 7ms difference between group A and group B, with a standard deviation of 20ms. This corresponds to a standardised mean difference of d = 0.35 (7 / 20). However, what would happen if we could measure our reaction times more precisely? Now instead of a standard deviation of 20ms, we have a more precise measurement of 7ms with a standard deviation of 10ms. This corresponds to standardised mean difference of d = 0.70 (7 / 10). This means we have doubled the standardised effect size by decreasing measurement error, but the unstandardised effect size has remained the same. This is an important point for designing experiments. Try and think carefully about how you are measuring your dependent variable. By using a more precise measure, you could decrease the number of participants you need while maintaining an adequate level of statistical power. It may not always be possible to half the variability, but even a 25% decrease in variability here could save you 114 participants in a between subjects design (two tailed, 80% power). To see how the

standardised effect size increases as you progressively decrease the variability in measurement, see this plot below:



As the standard deviation decreases, it makes it easier to detect the same difference between the two groups indicated by the decreasing red shaded area. For an overview of how measurement error can impact psychological research, see Schmidt & Hunter (1996).

If you are using a cognitive task, another way to decrease the variability is to increase the number of trials the participant completes (see Baker, 2019). The idea behind this is experiments may have high within-subject variance, or the variance of the condition is high for each participant. One way to decrease this is to increase the number of observations per condition, as it increases the precision of the estimate in each participant. Therefore, if you are limited in the number of participants you can collect, an alternative would be to make each participant complete a larger number of trials.

9. References

Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E., & Williams, M. (2020). Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences*, *38*(17), 1933–1935. https://doi.org/10.1080/02640414.2020.1776002

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*(2), 73–80. https://doi.org/10.1016/j.apergo.2004.01.002

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Bartlett, J. E., Jenks, R., & Wilson, N. (2020). *No Difference in Trait-Level Attentional Bias Between Daily and Non-Daily Smokers*. PsyArXiv. https://doi.org/10.31234/osf.io/cn64d

Bhogal, M. S., & Bartlett, J. E. (2020). Further support for the role of heroism in human mate choice. *Evolutionary Behavioral Sciences*. https://doi.org/10.1037/ebs0000230

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Coleman, T. J., Bartlett, J. E., Holcombe, J., Swanson, S. B., Atkinson, A. R., Silver, C., & Hood, R. (2019). Absorption, Mentalizing, and Mysticism: Sensing the Presence of the Divine. *Journal for the Cognitive Science of Religion*, *5*(1), 63–84. https://doi.org/10.31234/osf.io/k5fp8

Conti, A. A., McLean, L., Tolomeo, S., Steele, J. D., & Baldacchino, A. (2019). Chronic tobacco smoking and neuropsychological impairments: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *96*, 143–154. https://doi.org/10.1016/j.neubiorev.2018.11.017

Harms, C., Genau, H. A., Meschede, C., & Beauducel, A. (2018). Does it actually feel right? A replication attempt of the rounded price effect. *Royal Society Open Science*, *5*(4), 1–13. https://doi.org/10.1098/rsos.171127

Hobson, H. M., & Bishop, D. V. M. (2016). Mu suppression – A good measure of the human mirror neuron system? *Cortex*, *82*, 290–310. https://doi.org/10.1016/j.cortex.2016.03.019

Lakens, D. (2021). *Sample Size Justification*. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D., & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920951503. https://doi.org/10.1177/2515245920951503

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. https://github.com/richarddmorey/psychology_resolution/blob/master/paper/response.pdf

Neyman, J. (1977). Frequentist Probability and Frequentist Statistics. *Synthese*, *36*(1,), 97–131.

Pethrus, C.-M., Johansson, K., Neovius, K., Reutfors, J., Sundström, J., & Neovius, M. (2017). Suicide and all-cause mortality in Swedish deployed military veterans: A population-based matched cohort study. *BMJ Open*, *7*(9), e014034. https://doi.org/10.1136/bmjopen-2016-014034

Quintana, D. S. (2016). An effect size distribution analysis of heart rate variability studies: Recommendations for reporting the magnitude of group differences. *BioRxiv*, 072660. https://doi.org/10.1101/072660

Quintana, D. S. (2018). Revisiting non-significant effects of intranasal oxytocin using equivalence testing. *Psychoneuroendocrinology*, *87*, 127–130. https://doi.org/10.1016/j.psyneuen.2017.10.010

Rattan, A., Steele, J., & Ambady, N. (2019). Identical applicant but different outcomes: The impact of gender versus race salience in hiring. *Group Processes & Intergroup Relations*, *22*(1), 80–97. https://doi.org/10.1177/1368430217722035

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rode, J. B., & Ringel, M. M. (2019). Statistical Software Output in the Classroom: A Comparison of R and SPSS. *Teaching of Psychology*, *46*(4), 319–327. https://doi.org/10.1177/0098628319872605

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, *1*(2), 114–117. https://doi.org/10.1111/j.2041-210X.2010.00014.x

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological

Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*, 1–13. https://doi.org/10.3389/fpsyg.2019.00813

Schmeck, A., Mayer, R. E., Opfermann, M., Pfeiffer, V., & Leutner, D. (2014). Drawing pictures during learning from scientific text: Testing the generative drawing effect and the prognostic drawing effect. *Contemporary Educational Psychology*, *39*(4), 275–286. https://doi.org/10.1016/j.cedpsych.2014.07.003

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 8.

Seli, P., Risko, E. F., & Smilek, D. (2016). On the necessity of distinguishing between unintentional and intentional mind wandering. *Psychological Science*, *27*(5), 685–691. https://doi.org/10.1177/0956797616634068

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797